Limiting Freedom of Expression in Digital Platforms: Between State Obligations and Intermediary Moderation in a Multicultural Context

S M Morsalin Hider Ashik

Faculty of Law
McGill University

April 2022

A thesis submitted to McGill University in partial fulfillment of the degree of

LL.M. (Thesis)

Abstract

Digital platforms have reshaped our understanding of public communication. Anyone with smart gadgets with an internet connection can publish their opinion outside of our traditional understanding of media without any cost. These digital platforms provide opportunities to engage in public debate, no matter where anyone is. While promoting free expression worldwide, they apply their enforcement mechanism to make the internet safe for their users. Their enforcement mechanisms mostly remain publicly inaccessible, and at the same time, they are under tremendous pressure from governments to remove certain illegal and sometimes harmful content from their platforms. While doing so, their policies often resulted in the suppression of free expression around the globe. The transnational or rather global nature of digital platforms is creating challenges to determine what law is to be followed. Additionally, in the absence of any guiding principles in international human rights law for this 'special' type of media platform, it is also difficult to protect freedom of expression and other associated rights. This study examines the challenges of human rights-based content moderation by digital platforms aka intermediaries, focusing on the cultural diversity in the global platform. This study argues that existing international human rights laws' mechanisms in their present form are not adequate to address the complex legal challenges of digital platforms. I conclude by suggesting a need to have a universally accepted guiding principle where the limit of self-governance by intermediaries and platform governance by states will be balanced.

Limiter la liberté d'expression sur les plateformes digitales: entre obligations de l'Etat et modération par des intermédiaires dans un contexte multiculturel

Abstract

Les plateformes digitales contribuent à façonner notre compréhension de la communication publique. N'importe quelle personne avec un accès à des gadgets intelligents et une connexion internet peut désormais publier son opinion en dehors de tout media traditionnel et sans coût. Ces plateformes digitales fournissent des occasions pour le débat public, où que se trouvent les intervenants. Tout en tentant de promouvoir la liberté d'expression globalement, les plateformes digitales mettent en œuvre des mécanismes afin de rendre l'internet sécuritaire pour ses utilisateurs. Cependant, ces mécanismes de mise en œuvre demeurent largement inaccessibles au public, tout en subissant des pressions considérables de la part des gouvernements pour retirer certains contenus illégaux ou même simplement nocifs de leurs plateformes. Il en découle souvent une suppression effective de la liberté d'expression globalement. La nature transnationale ou même universelle des plateformes digitales crée des défis pour identifier le droit applicable. En outre, en l'absence de principes directeurs en matière de droit international des droits humains adaptés à ces types particuliers de plateforme médiatique, il est également difficile de protéger la liberté d'expression et les droits qui s'y rattachent. La présente thèse envisage les défis d'une modération par les plateformes digitales entendues comme intermédiaires en se concentrant sur la question de la diversité culturelle sur les plateformes globales. Elle soutient que les mécanismes existants de droit international de droits humains ne sont pas adéquats pour répondre aux défis juridiques complexes que rencontrent ces plateformes. Je conclue en suggérant le besoin d'avoir un principe universellement accepté qui limite l'auto-gouvernance des intermédiaires et la réconcilie avec la gouvernance des plateformes par les Etats.

Acknowledgment

First and foremost, I want to thank my supervisor, Professor Frédéric Mégret, whose constant guidance, insightful contribution, and encouragement were invaluable and this thesis would not have been possible without your support. Working with you is a lifelong experience. Thank you so much!

I want to thank the faculty and staff of the Faculty of Law for their continued support of the graduate law students. I also want to thank the Centre for Human Rights and Legal Pluralism and its members for their help and support. Thank you for all the help that you have provided me throughout my time at McGill.

I am grateful to Muhammad Rezaur Rahman, Alida Binte Saqi for their support and guidance in my McGill journey.

Last but not the least, I want to thank my family. Thank you to my parents for their constant faith in my ability.

CONTENTS

Chapter 1

Introduction

With the emergence of the internet, the primary medium of expression has fundamentally changed. It is no longer required to communicate with others in a different country simply by writing letters. Not even artists now rely only on traditional modes to reach their global audience. They can reach a global audience at a fingertip whilst sitting at home, with the help of the internet. Any expression posted on social media can reach people of various countries, cultures, races, religions, languages, etc. However, one person expressing an idea from his point of view may not be the same as the one from a different culture. There is a chance of misunderstanding expressions of thought in different contexts. This may even lead to cross-cultural conflict and/or inter-cultural conflict. One expression that may be appropriate for one religion may not be the same for others. One expression that might not be illegal in one country may be illegal in another.

The international law protecting freedom of speech and expression is not absolute. Article 19 of ICCPR imposes some restrictions on the exercise of the right. It states that in the exercise of the right, there shall be special duties and responsibilities which may be subject to certain restrictions, but these shall only be such as are 'provided by law' and are necessary: (a) for respect of the rights or reputations of others; (b) for the protection of 'national security' or of 'public order' or of 'public health' or 'morals'.[1] The scope of the restrictions imposed by Article 19(3) of ICCPR is not defined, and it is left open for the ratifying states to enact a law to impose such restrictions. Nonetheless, while enacting any law, states must follow the

[1] *International Covenant on Civil and Political Rights*, 19 December 1966, 999 UNTS 171 arts 9—14 (entered into force 23 March 1976) [*ICCPR*] Art. 19(3).

guideline prescribed in the article mentioned above, and such restrictions on expression must be subject to a strict test of *necessity* and *proportionality*.[2]

What is quite clear, however, is that the limits were designed with the state in mind and left some of the keywords undefined. They were not primarily designed, conversely, with a digital platform in mind. Of course, one might argue that freedom of expression is guaranteed irrespective of the media[3] of one's choice, and digital platform is merely one such media. However, we must not overlook the inherent impact and outreach of the kind of media through which expression manifests itself. A person posting from one place, using a platform incorporated in a particular place, will be subject to specific laws, but the content communicated will also be accessible from several places simultaneously by the receivers/audience. There is a sort of tri-angle involved in this whole system, with the state increasingly called upon to mediate freedom of expression at home and its impact abroad and *vice versa*.[4]

Nevertheless, when it comes to limiting harmful expression, states have struggled because of their limited sovereignty on digital platforms. To be sure, digital platforms have made our expression globally accessible in practice. Otherwise, there was little chance historically of expressing a view globally. Even if it was possible, that too was the subject of various forms of national censorship. For example: to release any Hollywood movie in various countries, it has to go through censorship in those countries. However, today anyone can watch a Hollywood movie from any part of the world without multiple levels of censorship by several

[2] Communication No. 1022/2001, *Velichkin v. Belarus*, (adopted on 20 October 2005).

[3] ICCPR, *supra* note 1 Art 19(2): Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or print, in the form of art, or through any other media of his choice.

[4] Jack M Balkin, "Free Speech is a Triangle" (2018) 118:7 Columbia L Rev 2011–2056.

governments. The state's ability to censor, and by extension to impose a certain unique sovereign reading of culturally inappropriate expression, is therefore weakened.

The very idea of global reach in digital platforms, then, is creating serious problems. Digital platforms are used by various groups with diverse agendas, be it the promotion of terrorism, pornography, cyberbullying, sexual harassment, "anti-national" activity, religious proselytism or denigration, etc. These issues are complicated when someone from one country is posting something on a digital platform that is measured through various lenses in different parts of the world. For example, a cartoon publication by Charlie Hebdo of Prophet Mohammad led to huge controversy.[5] In this case, the publication was legal in France, but it was bound to be interpreted differently when it was posted on a digital platform. As a result, the French government was forced to temporarily close embassies in 20 countries.[6] Another impact of such publication was a terrorist attack which caused the death of 14 people.[7] Another example is a highly controversial movie posted on YouTube titled 'Innocence of Muslims' that presented the Prophet Mohammad as a "foolish" and "power-hungry" man.[8] The impact of this controversial movie resulted in at least 75 deaths and more than 100 injured in different countries, including a US Ambassador.[9]

[5] Scott Sayare & Nicola Clark, "French Newspaper Publishes Cartoons Mocking Muhammad - The New York Times", *The New York Times* (19 September 2012), online: <https://www.nytimes.com/2012/09/20/world/europe/french-magazine-publishes-cartoons-mocking-muhammad.html>.

[6] *Ibid*.

[7] *Ibid*.

[8] "Q&A: Anti-Islam film", *BBC News* (20 September 2012), online: <https://www.bbc.com/news/world-middle-east-19606155>.

[9] See "Innocence of Muslims Controversy", *Berkley Center for Religion, Peace & World Affairs* (2013), http://berkleycenter.georgetown.edu/essays/em-innocence-ofmuslims-em-controversy>.

The problem, then, is one of translation: what may be legal in one part of the world may be very controversial in terms of culture and religion from other perspectives. A limited number of people engaged in their freedom of expression can have dire repercussions on others (innocents) who are subjected to various attacks in different parts of the world.

This creates dilemmas for the state. On the one hand, the state is trying to govern its cyberspace by using various strategies; on the other hand, the inherent design of cyberspace is creating a jurisdictional complexity. Moreover, digital platforms (hereinafter, intermediaries[10]) are also engaged in governing expressions posted on their respective platforms. Which one should prevail in case of conflict between the platform's policy and state law? When can a state intervene in the platform's decision? What will be the guiding principle of platform governance both for the platform owners and states? Is there any guidance in International Human Rights Law (IHRL)?

This thesis addresses the existing IHRL's appropriateness to govern expressions in digital platforms as well as self-governance by intermediaries while dealing with freedom of expression. Special attention is given to the multicultural and global digital platform while addressing these issues. The thesis' central research question, then, is: How can IHRL strike a balance between self-governance and state intervention in a digital platform operating as part of a transnational and multicultural environment?

[10] Intermediary generally means the bridge between the author and the audience. Intermediaries are labelled as "publishers" and "distributors," or "publishers" and "secondary publishers." In this thesis, Intermediaries are referred to the digital platforms. See for more details: Christina Mulligan, "Technological Intermediaries and Freedom of the Press" (2013) 66 SMU L Rev 157; Felix T. Wu, "Collateral Censorship and the Limits of Intermediary Immunity" (2011) 87 Notre Dame L Rev 293; Danielle Keats Citron & Helen Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age" (2011) 91 BUL Rev 1435; Balkin, *supra* note 4; Jack M Balkin, "The Future of Free Expression in a Digital Age" (2009) 36:2 Pepp L Rev 427.

To this end, the thesis will examine existing IHRL governing freedom of expression and the nature and impact of these rules on digital platforms. It is essential to discuss the nature of the digital platform and its characteristics. Without a proper understanding of the characteristics of this relatively new media, it will not be possible to govern expression effectively.

Later on, this thesis will analyze the self-governance of intermediaries (content moderation)[11] , with particular attention to how they interpret their cultural context. The inherent global nature of this new media platform deals with content uploaded by various users from different cultures. It is crucial to consider the cultural context while making any content decision. At the same time, the thesis will examine the online censoring mechanism of private platforms, the outcomes of such mechanisms, and their impact.

This research adopts a largely doctrinal method to analyze the core IHRL provisions (in particular, provisions relating to freedom of expression and provisions relating to other related rights) to determine whether IHRL in its present form is suited to new media platforms in governing freedom of expression. Although the thesis does not pursue any comparative analysis per se, it will refer to the various laws and regulations of platforms governance by different countries and regions to illustrate and explain particular concepts. It will rely on both primary and secondary sources in terms of collecting data. The primary sources are international treaties and conventions, UN reports, legislations and regulations (domestic and regional), and case laws. The secondary sources include books, journals, platforms policies, newspaper reports, and scholarly writings on a similar topic.

This research is timely and addresses a range of crucial issues. Firstly, the question of inter-cultural moderation is largely unresolved in the sense that a still significant debate exists around its nature and form. Secondly, the existing literature on this field of study mainly deals with

[11] Roberts defines content moderation as, "the organized practice of screening User-Generated content (UGC) posted to internet sites, social media, and other online outlets. Sarah T. Roberts, Behind The Screen: The Hidden Digital Labor of Commercial Content Moderation, (2014) Phd Thesis, University of Illinois, Chicago, IL

the topic broadly relating to self-governance by intermediaries[12], the question of liability[13], application of human rights in content moderation[14], limitation of international law[15], and regulation of digital media by states.[16] My thesis, by contrast, focus on the overall governance of speech both by the intermediaries and states. Additionally, the main focal point of this thesis is the cultural aspect of a global platform. Cultural context/variables are often overlooked or inadequately explored in the scholarships. Thirdly, in recent years governments have been taking various steps to govern digital platforms, and those are worth addressing in themselves as a response that may help alleviate the problem but also has the potential to make things worse. Finally, platforms are coming up with updated enforcement mechanisms to regulate their content regularly, and those need to be addressed. This is why I believe this thesis will significantly contribute to a better understanding of this complex field.

The thesis is divided into four chapters including the introductory one. Chapter two discusses the nature and impact of any expression expressed on digital platforms and the unique characteristics of digital platforms. It analyzes the existing human rights laws governing

[12] Balkin, *supra* note 4; Angelo Jr Golia, "Beyond Oversight: Advancing Societal Constitutionalism in the Age of Surveillance Capitalism" (2021) SSRN Journal, online: <https://www.ssrn.com/abstract=3793219>; Sarah Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms" (2018) 20:11 New Media & Society 4366–4383; Sten Schaumburg-Müller, "Private Life, Freedom of Expression and the Role of Transnational Digital Platforms: A European Perspective" in YSEC Yearbook of Socio-Economic Constitutions (Cham: Springer International Publishing, 2022); Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech" (2018) 131 Harv L Rev 1598–1670.

[13] Hayden Benge, *Who's liable? The Intersection of Free Speech and Content Regulation on Social Media Platforms* (Honors Thesis, UniversityofMississippi, 2019) [unpublished]; Natali Helberger, Jo Pierson & Thomas Poell, "Governing online platforms: From contested to cooperative responsibility" (2018) 34:1 The Information Society 1–14; Mark Bunting, "From Editorial Obligation to Procedural Accountability: New Policy Approaches to Online Content in the Era of Information Intermediaries" (2018) 3:2 J Cyber Poly, online: <https://www.ssrn.com/abstract=3185005>.

[14] Thiago Dias Oliva, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression" (2020) 20:4 H R L Rev 607–640; Barrie Sander, "FREEDOM OF EXPRESSION IN THE AGE OF ONLINE PLATFORMS": (2020) 43:4 Fordham Intl L J 68.

[15] Evelyn Douek, "The Limits of International Law in Content Moderation" (2021) 6 UC Irvine J Intl Transnational & Compa L, online: <https://www.ssrn.com/abstract=3709566>.

[16] "Free Speech, Media Freedom and Regulation of Online Speech" in *Dimensions of Free Speech* Philosophy and Politics - Critical Explorations (Cham: Springer International Publishing, 2021) 93.

freedom of expression. It explores the key historical events to illustrate the development of international law on the matter. Drawing on history and existing legal developments, it argues that existing IHRL in its present form cannot govern freedom of expression on the digital platform.

Following on from the argument set out in chapter two, chapter three addresses content moderation by intermediaries and the state's attempt to govern digital space. It focuses mainly on the cultural aspects of content and discusses cultural variables with examples from both ends: intermediaries and users. It also acknowledges some recent development by one of the intermediaries to accommodate cultural representation and diversity in their content moderation system. Furthermore, it provides a brief discussion on the liability of intermediaries. The thesis concludes that since intermediaries' liability is not established globally, and in the absence of any universally recognized guiding principle for content moderation, neither the state nor intermediaries can protect freedom of expression globally.

Chapter four sums up the whole thesis. I conclude by suggesting a need to have a universally accepted guiding principle where the limit of self-governance by intermediaries and platform governance by states will be balanced.

Chapter 2

Regulating ‘Modern Platform’ Using the ‘Old’ Legal Framework: An Analysis

2.1 Introduction

The emergence of internet-based social and media platforms came as a blessing for various reasons. At the same time, it has raised serious concerns regarding its governance internationally. When it comes to regulating expressions from a human rights standpoint, only a few provisions in international human rights law were developed more than a half-century ago. These relatively old legal frameworks are still being applied to regulate the relatively new type of media platform. This chapter argues that these legal frameworks are not adequate to regulate this modern platform.

2.2 Nature and Impact of Expressions on the Internet

Thousands of people express their views, ideas, contents, videos, etc., on the Internet throughout the world. Every individual is a content creator in this age of the Internet. The Internet provides the opportunity to express oneself of being in their comfort zone. A person posting his views on the Internet might not express the same on offline media. The Internet provides an option to remain distant from one's audience. Remaining beyond the reach of the audience makes one feel comfortable. The shy person gets the joy of expressing his thought on the Internet without fear. It helps him to boost his confidence. On the contrary, a person with an ill motive uses the Internet to fulfill his evil deeds.

Among all the internet platforms, social media platforms are the most impactful. Platforms such as Facebook, Twitter, and Instagram are the most influential ones. These platforms play

crucial roles in political publicity,[17] election results,[18] and terrorist propaganda[19] to name a few. Any content on these platforms is designed to share/retweet or comment,[20] using hashtags will make content easily accessible, videos go viral within a very short period of time,[21] and trolling culture on these platforms is becoming a serious issue.[22] Online content has the ability to deceive people on unprecedented scales,[23] encourage them to participate in enormous, coordinated fundraisers,[24] and mobilize the masses to perform strange behavior such as pouring buckets of ice water on their heads.[25] All of this takes place on algorithmic digital platforms whose data and ad-driven business model–called "surveillance capitalism" by Shoshana Zuboff–is explicitly geared to emphasize viral and emotive material that grabs our attention.[26] It has been demonstrated that profit-driven social media algorithms may control our access to information, provoke certain emotional responses, and even impact our emotions. Due to widespread and systematic access to digital media, speech may now travel faster than ever before, affecting listeners in unprecedented ways.

---

[17] Kevin Roose, Political Donors Put Their Money Where the Memes Are, *New York Times* (7 August 2017), online <www.nytimes.com/2017/08/06/business/media/politicaldonors-put-their-money-where-the-memes-are.html>

[18] Vindu Goel, "In India, Facebook's WhatsApp Plays Central Role in Elections" *The New York Times* (14 May 2018), online: <www.nytimes.com/2018/05/14/technology/whatsapp-india-elections.html>.

[19] Alexander Tsesis, "Terrorist Speech on Social Media", (2017) 70 Vand L Rev 651.

[20] Shea Bennett, 10 Easy Ways to Get More Retweets on #Twitter, (5 January 2015), online: *Adweek*, <www.adweek.com/digital/get-more-retweets-twitter/>.

[21] Ilya Pozin, 6 Qualities to Make Your Videos Go Viral, (7 August 2014) online: *Forbes* <www.forbes.com/sites/ilyapozin/2014/08/07/6-qualities-to-make-yourvideos-go-viral>.

[22] Peter Suciu, "Trolls Continue To Be A Problem On Social Media", online: *Forbes* <https://www.forbes.com/sites/petersuciu/2020/06/04/trolls-continue-to-be-a-problem-on-social-media/>.

[23] Michael Edmund O'Neill, Old Crimes in New Bottles: Sanctioning Cybercrime, (2000) 9 Geo. Mason L. Rev 237.

[24] Enrique Estelles-Arolas & Fernando Gonzales-Ladron-de-Guevara, "Towards an Integrated Crowdsourcing Definition", (2012) 38 J Info Sci 189 at 197 ("crowdsourcing" as "a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task").

[25] Kathy Giusti, "The Real Ice Bucket Challenge", online: *Time* <https://time.com/3204261/the-real-ice-bucket-challenge/> (explaining how the social media "ice bucket challenge" started with a single email to 60,000 people and that it increased donations for research into ALS (amyotrophic lateral sclerosis) to $94.3 million from the $2.7 million that had been raised in the same period the previous year).

[26] Shoshana Zuboff, *The Age of Surveillance Capitalism*, (London: Profile Books, 2019).

The long-standing geopolitical disagreements over cyberspace governance have created a structural and normative vacuum in the absence of universal international rules, which are progressively being filled by a small group of private non-state entities. Internet mega-platforms are informally and slowly adopting the role of international lawmakers in regulating online speech, wielding a historically unprecedented level of power over the public sphere. At the same time, increasing exposure to terrible governance failures raises severe problems about profit-seeking technology corporations' competence and legitimacy in pioneering internet governance–a phenomenon known in political science as 'norm entrepreneurship.[27] Mega internet companies have come to function as both lawmakers and judges of byzantine corporate laws on online content moderation, driven by opaque algorithms and advertisement-based business models that drive spectacular (and at times dangerous) virality.[28]

Probably most significantly, UN investigators concluded in March 2018 that Facebook had a "determining role" in a campaign of crimes against Myanmar's Rohingya Muslims that was defined as a "textbook example of ethnic cleansing"[29] with "hallmarks of genocide".[30] After a few months, the UN Fact-Finding Mission in Myanmar issued a report calling for an independent investigation into Facebook's role in inciting offline violence, claiming that the social media platform had been "a useful instrument for those seeking to spread hate" and that its response had been "slow and ineffective".[31] The New York Times conducted a

[27] Martha Finnemore & Kathryn Sikkink, "International Norm Dynamics and Political Change" (1998) 52:4 Intl Organizations 887.

[28] Kaya Yurieff, "Facebook's 'supreme court' just ruled against Facebook", *CNN* (28 January 2021), online: <https://www.cnn.com/2021/01/28/tech/facebook-oversight-board-first-decisions/index.html>.

[29] High Commissioner for Human Rights, Opening Statement to the 36th session of the Human Rights Council, 11 September 2011.

[30] GA, Report of the Special Rapporteur on the situation of human rights in Myanmar, Advance Unedited Version, A/HRC/37/70, 9 March 2018, at para. 65.

[31] Human Rights Council, "Report of the independent international fact-finding mission on Myanmar", 12 September 2018, A/HRC/39/64 at para 74 [UN FFM Report]. The report recommended that several public officials including a Senior-General Min Aung Hlaing with 2.9 million Facebook followers be prosecuted for spreading hate speech. Similarly, according to a recent Reuters investigation, Facebook has spent very little time and money trying to control hate speech in Myanmar over the years. Only 60 individuals were assessing

groundbreaking investigation shortly after that found that members of the Burmese Tatmadaw, reportedly numbered over 700 people, were the main operatives behind a sophisticated anti-Rohingya social media campaign that extended half a decade.[32] The prevalence of unfettered online hate speech was judged a vital component in motivating and legitimizing the atrocities perpetrated against the Rohingya in a relatively isolated society new to the Internet and afflicted by a "crisis of digital literacy".[33]

Incidents where online digital platforms are being used as a tool to violate human rights are alarming in nature. Online jihadist hate speech is said to have played a key role in radicalizing adolescents and instigating independent terror acts against civilians in Syria and Iraq.[34] Misinformation shared via Facebook and WhatsApp in Sri Lanka prompted widespread and violent anti-Muslim riots, forcing the government to shut down social media networks for weeks.[35] In South Sudan, inflammatory digital sharing is stoking the fires of tribal and ethnic conflict that threatens to escalate into genocide.[36] According to a 2016 UN report, "[S]ocial media has been used by partisans on all sides, including some senior government officials, to

allegations of hate speech and other information uploaded by Myanmar's 18 million active Facebook users as of April 2018: Steve Stecklow, "Why Facebook is losing the war on hate speech in Myanmar" (15 August 2018) *Reuters*, online: < https://www.reuters.com/investigates/specialreport/myanmar-facebook-hate/ >.

[32] Paul Mozur, "A Genocide Incited on Facebook, With Posts From Myanmar's Military" (Oct. 15, 2018) *The New York Times*, online:< https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>.

[33] An assessment report commissioned by Facebook found that "[a] large population of internet users lacks the basic understanding of how to use a browser, how to set up an email address and access an email account, and how to navigate and make judgments on online content. Despite this, most mobile phones sold in the country come preinstalled with Facebook." See Business for Social Responsibility, "Human Rights Impact Assessment: Facebook in Myanmar" (2018) at 12, online (pdf):< https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf >.

[34] See, Tom De Smedt, Guy de Pauw, Pieter Van Ostaeyen, "Automatic Detection of Online Jihadist Hate Speech" (February 2018) CTRS-007 at 3; Robert S. Tanenbaum, "Preaching Terror: Free Speech or Wartime Incitement," (2005) 55 American U L Rev 785; Jytte Klausen et al., "The YouTube Jihadists: A Social Network Analysis of Al-Muhajiroun's Propaganda Campaign," Perspectives on Terrorism 6, no. 1 (2012).

[35] Michael Safi, "Sri Lanka accuses Facebook over hate speech after deadly riots" (14 March, 2018) *The Guardian*, online:< https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hatespeech>.

[36] Justin Lynch, "In South Sudan, Fake News Has Deadly Consequences" (09 June, 2017) *The Slate*, online:<http://www.slate.com/articles/technology/future_tense/2017/06/in_south_sudan_fake_news_has_deadly_consequences.html>.

exaggerate incidents, spread falsehoods and veiled threats, or post outright messages of incitement."[37] This is similar to situations in Bangladesh,[38] Cambodia,[39] the Central African Republic,[40] Cameroon,[41] India,[42] and the Philippines[43] where social media platforms are similarly used as a tool by both states and non-state actors to incite communal tensions, with false news having as stated by Facebook "life or death consequences".[44]

2.3 The Historical Role of Media in the Development of International Law

Historically, media has played a direct role in several human rights atrocities and a passive role in developing international human rights law. Media campaigns aimed at inciting hatred, in which political or religious leaders incite the sentiments of would-be perpetrators, are frequently used to foreshadow international crimes. Commentators highlighted the role of the media, especially *Radio Télévision Libre des Milles Collines* (RTLM) and the *Kangura* magazine, in inciting acts of violence against the Tutsi minority in the 1994 Rwanda Genocide,

[37] "Letter dated 15 November 2016 from the Panel of Experts on South Sudan", UN Doc off S/2016/963, para 24.
[38] Kris Thomas, "6 Deaths, 450 Arrests and Mass Protests. It Started With a Facebook Post.", *VICE* (20 October 2021), online: <https://www.vice.com/en/article/y3vmpb/bangladesh-violence-facebook-post>.
[39] Mathew Ingram, "In some countries, fake news on Facebook is a matter of life and death" (21 November 2017) *Columbia Journalism Review*, online :< https://www.cjr.org/analysis/facebook-rohingya-myanmar-fake-news.php >.
[40] Lisa Schlein, "Hate Speech on Social Media Inflaming Divisions in CAR" (02 June 2018) *VOA News*, online: <https://www.voanews.com/a/hate-speech-on-social-media-is-inflaming-divisions-in-centralafricanrepublic/4420555.html>.
[41] "Burning Cameroon: Images you're not meant to see" (25 June 2018) *BBC News* (Prime Minister Philemon Yang has blamed Cameroonians living abroad for using social media to "spread hate speech and terror" and "order murders"), online: < https://www.bbc.com/news/world-africa-44561929>.
[42] "Social media rumors trigger violence in India; 3 killed by mobs" (May 25 2018) *NBC News*, online :<https://www.nbcnews.com/news/world/social-media-rumors-trigger-violence-india-3-killed-mobs-n877401>; Lauren Frayer, "How the Spread of Fake Stories in India Has Led to Violence" (July 17 2018) *NPR*, online: <https://www.npr.org/2018/07/17/629896525/how-the-spread-of-fake-stories-in-india-has-led-to-violence>.
[43] Lauren Etter, "What Happens When the Government Uses Facebook as a Weapon?" (7 December 2017) *Bloomberg Businessweek*, online: < https://www.bloomberg.com/news/features/2017-12-07/how-rodrigo-duterteturned-facebook-into-a-weapon-with-a-little-help-from-facebook>; Mathew Ingram, "Facebook now linked to violence in the Philippines, Libya, Germany, Myanmar, and India" (5 September 2018) *Columbia Journalism Review*, online: <https://www.cjr.org/the_media_today/facebook-linked-to-violence.php>.
[44] Sara Su, "Update on Myanmar" (15 August, 2018) *Facebook Newsroom*, online: <https://newsroom.fb.com/news/2018/08/update-on-myanmar/>.

with some advancing a theory of "radio genocide" and "death by radio"[45] and others observing that the primary weapons of the genocide were "the radio and the machete."[46] Unfortunately, the link between media propaganda and mass violence was not unique to the Rwandan situation. Around 80 years before, the Young Turk propaganda weekly '*Harb Mecmuasi*' disseminated propaganda material encouraging support for the genocide of 1.5 million Armenians, in part by convincing Turks of "the need to 'rid ourselves of these Armenian parasites'" and identifying them with "traditionally unclean animals such as rats, dogs, and pigs."[47] Roughly after two decades, the Nazi propaganda outlets such as the monthly '*Der Stürmer*' used the Turkish model to gradually mobilize support for the slaughter of six million Jews, eventually advocating for their annihilation "root and branch."[48] *The Reichsministerium für Volksaufklärung und Propaganda* (Reich Ministry of Public Enlightenment and Propaganda) of Joseph Goebbels seized all means of communication in Germany, including the press, music, cinema, and theatre, to advance a sophisticated and pervasive media campaign depicting Jews "as disease-carrying insects or vermin, tumors/tuberculosis that infected healthy Germans and thus had to be exterminated."[49] Similarly, observers emphasized the role of inciting speech by the Serbian Democratic Party (SDS) in igniting Bosnian-Serb violence

[45] Alan Thompson, *The Media and the Rwanda Genocide*, (London/Ottawa: Pluto Press/International Development Research Centre, 2007) at 165.

[46] Frans Viljoen, "Inciting violence and propagating hate through the Media: Rwanda and the limits of international criminal law" (January 2005) 26:1 Obiter 58.

[47] David Livingstone Smith, *less than Human: Why we Demean, Enslave and Exterminate Others* (New York: St. Martin's Press, 2011) at 145.

[48] Erin Steuter & Deborah Wills, *At War with Metaphor: Media, Propaganda, and Racism in the War on Terror* (Lanham: Lexington, 2009) at 142.

[49] See Gregory S Gordon, Atrocity Speech Law: Foundation, Fragmentation, Fruition (Oxford: Oxford University Press, 2017) see Chapter 1 "Speech and Atrocity: An Historical Sketch" at 30.

against Muslims in the post-Cold Balkan War. Among them, some of the observers claimed that "[e]veryone killed in this [Bosnian] war was killed first in the newsroom."[50]

These historical events in the 20th century are the darkest examples of using media campaigns to incite hatred and violence that resulted in mass atrocities. Media campaigns are carried out both in written and oral form to disseminate speech. Moreover, international law has recognized it in international criminal law and international human rights law. Despite his lack of direct engagement with the military or the Holocaust, Nazi propagandist and '*Der Stürmer*' owner Julius Streicher was convicted of crimes against humanity and hanged by the International Military Tribunal Nuremberg.[51] Hans Fritzsche, chief of Goebbel's Propaganda Ministry's Radio Department, was also charged with crimes against humanity though eventually acquitted.[52] Also, the Nazi Press Chief Otto Dietrich was convicted of crimes against humanity under Control Council No. 10 for his "well-planned, often repeated, continuous effort to incite hatred of the German people towards Jews".[53] Following the second world war, the Convention on the Prevention and Punishment of the Crime of Genocide,1948 (mostly referred to as the Genocide Convention) created the international crime of "direct and public incitement to commit genocide" in Article III(c), overriding the American delegation's rejection based on free speech concerns.[54] Similarly, the European Convention on Human Rights (1950) recognized limitations on freedom of expression as "prescribed by law and [that] are necessary for a democratic society, in the interests of national security, territorial integrity, or public

[50] John Oppenheim & Willem-Jan Van der Wolf, *Global War Crimes Tribunal Collection* (Nijmegen, The Netherlands: Global Law Association, 1997) at 148; see also *Prosecutor v Brđanin*, IT-99-36-T, Judgment, para 80 (ICTY, Sept. 1, 2004).
[51] *Ibid.*
[52] *Ibid* at 186–187.
[53] *United States v. von Weizsaecker*, Judgment (Intl Mil Trib, Sept 30, 1946), *reprinted in* F.R.D. 161–163 (1946).
[54] Richard Ashby Wilson, *Incitement on Trial: Prosecuting International Speech Crimes*, (Cambridge: Cambridge University Press, 2017) in Cambridge Studies in Law and Society.

safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others."[55]

A similar provision was introduced by the Convention on the Elimination of All Forms of Racial Discrimination (1965), in its Article 4(1) stating "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another color or ethnic origin[...]" and stated that States Parties "[s]hall not permit public authorities or public institutions, national or local, to promote or incite racial discrimination."[56] One of the most widely ratified international conventions, ICCPR puts some restrictions on the freedom of expression and of speech "[f]or respect of the rights or reputations of others; [f]or the protection of national security or of public order; or of public health or morals."[57] Additionally, Article 20 prohibits "[a]ny propaganda for war ...; [a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence [...]." The Committee on the Elimination of Racial Discrimination and the Human Rights Committee, both quasi-judicial authorities that monitor states' compliance with their respective human rights treaties, have issued recommendations attempting to define the scope of these duties.[58]

These international instruments work as a framework for protecting human rights all over the world. However, these treaties also empower states to enact their law and define the scope of

[55] Council of Europe, Convention for the Protection of Human Rights and Fundamental Freedoms, Rome, 4.XI.1950.

[56] *International Convention on the Elimination of All Forms of Racial Discrimination,* 21 December 1965, 660 UNTS 195 Art. 4(1) (entered into force 4 January 1969). [ICERD]

[57] ICCPR, *supra* note 1, Art 19 (3).

[58] Human Rights Committee: *J.R.T. and the W.G. Party v Canada*, Communication No 104/1981 (18 July 1981), UN Doc A/38/40 (Supp No. 40) at 231; *Kasem Said Ahmad and Asmaa Abdol-Hamid v Denmark*, Communication No 1487/2006, UN Doc CCPR/C/92/D/1487/2006, 18 April 2008; *Malcolm Ross v Canada*, Communication No. 736/1997 (1 May 1996), UN Doc CCPR/C/70/D/736/1997, 18 October 200; Committee on the Elimination of Racial Discrimination: *L.K. v The Netherlands*, Communication No 4/1991, UN Doc CERD/C/42/D/4/1991, 16 March 1993; *L.R. v Slovak Republic*, Communication No 31/2003, UN Doc CERD/C/66/D/31/2003, 10 March 2005; *Quereshi v Denmark*, Communication No 33/2002, UN Doc CERD/C/66/D/33/2003, 10 March 2004.

the restrictions. The general comment on the freedom of opinion and expression provides some guidelines for states while imposing restrictions on the right by enacting any law. It suggests that the restrictions "must not be overbroad"[59] and a strict "test of necessity and proportionality"[60] will be applicable. Despite such clarification, states are reluctant to apply those guidelines in ways that do not permit the enjoyment of the right to its fullest.[61] However, states follow the guidelines to govern offline speech in their respective territories.

Nevertheless, when it comes to governing online speech and expression, states are applying the same old principles in online speech governance due to the lack of international law governing online speech. Before going to the legal aspects of online speech governance by applying old laws, it is crucial to understand the unique features of online speech on digital platforms. Understanding the uniqueness of this "special media" platform will be possible to develop laws to govern them. The following section discusses the particular characteristics of digital platforms.

2.4 Unique Features of Expressions in the Digital Platforms

As argued before, expression in "traditional media"[62] is not the same as an expression on digital platforms. Without understanding the uniqueness of this relatively new media platform, it is not possible to regulate it efficiently. It is worth underlining that digital platforms are different

[59] *General comment no. 34, Article 19, Freedoms of opinion and expression*, UN Human Rights Committee (HRC), 102nd Session CCPR/C/GC/34, 12 September 2011 at para 34.

[60] *Ibid* at paras 22, 33.

[61] Some of the few states that have raised reservations about Article 20 do not even oppose the article itself, but rather declare that they have previously passed relevant legislation and reserve the right not to legislate further on the subject. In fact, the United States has included a reservation noting that "article 20 does not permit or require legislation or other action by the United States that would impair the right to free expression and association protected by the Constitution and laws of the United States." See, United Nations Treaty Collection, Chapter IV, Human Rights, ICCPR, New York, 16 December 1976, online: <https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATYandmtdsg_no=IV4andchapter=4andlang=en#ENdDec>.

[62] Traditional media existed before the invention of the Internet. Examples of such media are newspapers, magazines, TV, radio, movie, any form of art, lift lets etc.

from traditional media in terms of accessibility, outreach, size of audiences, perpetuity, or secrecy. Their impact also varies at different levels in the online platforms.[63] Actually, what does that mean?

a. Cyber-Psychology and Secrecy

The psychology of the digital media platforms user works differently than those who do not use those media. There is substantial social science data that suggests that the Internet's facilitation of anonymous users and anonymous communication (or perceptions thereof) emboldens individuals to be more hateful than they would otherwise be.[64] This cyber-psychological phenomenon is based on a sense of liberty from traditional standards of behaviour, as well as a sense of impunity, which motivates online speakers to their worst thoughts and actions.[65] In an experiment conducted by Philip Zimbardo called 'Stanford prison experiment,' it was demonstrated that secrecy in groups can lead to progressively violent and even cruel behavior.[66] And in relation to group dynamics, it is found that social media fosters an 'illusion of huge number'—for example, the number of times a post has been 'liked,' 'shared,' or 'retweeted' that encourages users to overestimate how many people share their

[63] Alexander Brown, "What is so special about online (as compared to offline) hate speech?" (2018) 18: 3 Ethnicities at 297; Citron DK, *Hate Crimes in Cyberspace*, (Harvard, MA: Harvard University Press, 2014); Citron DK and Norton H, "Intermediaries and hate speech: Fostering digital citizenship for our information age" (2011) 91: Boston U L Rev 1435.

[64] *Ibid* Brown at 298.

[65] Daniel J Solove, *The Future of Reputation: Gossip, Rumor, and the Privacy on the Internet*, (London: Yale University Press, 2007) at 17; Citron (2014), *supra* note 63 at 57, 59–60.

[66] Philip Zimbardo "*The Lucifer Effect: Understanding How Good People Turn Evil*" (New York City: New York Random House, 2013).

viewpoint or in other words 'like-minded'.[67] This raises the risk of "confirmation bias," which confirms prior beliefs, which may legitimate and encourage otherwise fringe hate speech.[68]

Additionally, according to several studies, the Internet makes social interactions more "asynchronous."[69] Such that, although it allows for direct and fast transmission, it also allows for large delays between conversations. The flexibility to go in and out of a discussion, in essence, supports "conversational relaxation," which permits huge groups of individuals to connect for extended periods of time.[70] However, it also implies that these groups are not subjected to the instant reactions/feedbacks of their recipients. According to psychologist John Suler, the absence of "a continuous feedback loop that encourages certain behaviors and extinguishes others" generated by asynchronous communication causes an "online disinhibition effect" that can lead to particularly violent groups. Furthermore, others argue that the lack of nonverbal indications from the audience encourages unconstrained and perhaps abusive speech.[71]

b. Perpetuity, Immediacy, and Itinerancy

The rapid posting features of digital platforms (as opposed to offline mediums such as printed flyers, public speeches, posters, newspapers, and so on) can inspire hate speech that is unfiltered, impulsive, and uncontrolled. Anything posted on the digital platform remains

[67] Katelyn YA McKenna and John A Bargh, Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology (2000) 4 Personality & Soc Psychology Rev 57 at 64.

[68] Karsten Muller and Carlo Schwarz, "Fanning the Flames of Hate: Social Media and Hate Crime" Working Paper Series, (2018) at 2, *University of Warwic*k online (pdf): <https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/373-2018_schwarz.pdf>.

[69] Therese Enarsson & Simon Lindgren, "Free speech or hate speech? A legal analysis of the discourse about Roma on Twitter" (2018) Information & Communications Technology L at 4.

[70] Joseph B. Walther, "Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction" (1996) 23: 1 Communication Research 3 at 26.

[71] M.J Culnan & M-L Markus, 'Information Technologies' in L. Putnam & D. Mumby (Eds.), Handbook of Organizational Communication (Beverly Hills, CA: Sage 1987) 420 at 429.

forever (in one form or another by the archive, screenshot, etc.). It becomes available immediately to its targeted audience. As Alexander Brown explains:

> On the Internet, the time difference between having a thought or sentiment and expressing it to a specific individual located a considerable distance away, a group of like-minded individuals, or a large audience can be within seconds. In contrast, if a general public member wishes to communicate a group libel against Jews to a large number of people using traditional media, it can take a significant amount of effort to create and print leaflets and distribute them out on the street, or mail them to individuals. It also requires time to generate an automated mobile message, set up the required phone accounts, collect a list of phone numbers, and execute the automated calls. The point is that the Internet allows for and promotes fast reactions that are, by definition, more spontaneous in a sense described above.[72]

Undoubtedly, the Internet has significantly changed the circumstances and dynamics of social communication. Moreover, since online speech is fast, the law must pay special attention to its distinguishing features to be effective. As Marshall McLuhan commented, "the medium is the message,"[73] Social media's unique features should be considered while evaluating the applicability of human rights norms. In this regard, it is worth mentioning that considering the nature of the medium of communication in such assessments is not completely without a jurisprudential basis.[74] For example, in *Arslan v. Turkey* (cited with acknowledgment by the ICTR in the Media case) the European Court of Human Rights (ECtHR) overturned the conviction of an award-winning journalist whose book, *History in Mourning, 33 Bullets,*

[72] Brown, *supra* note 63.

[73] Marshall McLuhan, *Understanding Media: The Extensions of Man,* (New York City: McGraw-Hill, 1964) chapter 1.

[74] See Gordon, *supra* note 49, suggesting "channel of communication" as a relevant criterion for incitement to genocide; at 299.

portrayed Turks as barbarous invaders who massacred Kurdish families.[75] Carol Pauli writes about the Court's assessment of media type:

> The [ECtHR in *Arslan v Turkey*] was seemingly more tolerant of a book than it would have been of other modes of communication. It discovered that literary works were less likely to affect national security and public order than mass media (probably meaning broadcast media).[76]

Overall, online speech's unique characteristics and dynamics–anonymity/secrecy, perpetuity, invisibility, immediacy, openness and accessibility, communitarian, and sometimes libertarian attitude allow for inexpensive and fast mass communication and virtual content transmission. The Internet's trans-nationality, which allows for borderless publishing and a vast audience, broadens the reach of online hate speech in terms of the harm done by "ramping up the public humiliation factor."[77]

Are international human rights laws (IHRL) capable of resolving the legal concerns raised by the unique features of online media platforms, the type of content published on those platforms, and violence? If not, what are the challenges with the existing mechanisms?

2.5 International Human Rights Law Governing Freedom of Speech and Expression

As previously mentioned, the idea of private non-state actors capable of conducting effective hate campaigns outside of the government machinery was essentially unanticipated by the drafters of the key international human rights instruments.[78] However, those standards are still

[75] *Arslan v. Turkey*, (Application no. 23462/94) Eur Ct HR (1999)

[76] *Ibid; See* also, Carol Pauli, Killing the Microphone: When Broadcast Freedom Should Yield to Genocide Prevention, (2010) 61: 4 Ala L Rev 665 at 686.

[77] Brown, *supra* note 63 at 307.

[78] International human rights responsibilities prohibiting hate speech were developed in an environment where states were the primary actors with the resources to generate, and more precisely, spread, such content. According to George Gordon, the traumatic historical experiences that influenced the drafting of the UDHR, the ICCPR, and subsequent international treaties mainly involved state-sponsored hate propaganda carried out with the full weight

being applied to regulate digital platforms. This section argues that the existing laws cannot address the incitement on digital platforms.

The modern terrorist groups to commit heinous crimes in the 21st century have access to reach a global and large audience with the help of Facebook, Twitter, and other digital platforms.[79] For example, at its peak in late 2014, the Islamic State's (IS) skilled propaganda wing operated 46,000 social networking sites accounts,[80] generating a staggering 90,000 posts each day across

of the government apparatus, particularly the Armenian Genocide and the Holocaust. See Gordon, *supra* note 50 at 30. The same is also seen by the shift in vocabulary from hate propaganda to hate speech. When forming a Special Committee on Hate Propaganda in1967, for example, the Canadian government was still using the former. David McGoldrick & Therese O'Donnell, "Hate-speech laws: consistency with national and international human rights law" (1998) 18:4 Legal Studies 453 at 459. See also Jona Adalheidur Palmadottir & Iuliana Kalenikova, "Hate speech; an overview and recommendations for combating it" (2018) *Icelandic Human Rights Centre* online (pdf): <http://www.humanrights.is/static/files/Skyrslur/Hatursraeda/hatursraedautdrattur.pdf> at 4: "There is a distinction to be made between hate speech and hate propaganda. Hate propaganda is systematic and frequently refers to a certain ideology, such as Nazi Germany's hate propaganda against Jews. Hate speech is uttered by various individuals who do not necessarily know one another and is hence not systematic."

[79] See for e.g. Thomas Tracy, "ISIS has mastered social media, recruiting 'lone wolf' attacks to target Times Square: Bratton", online: *NY Daily News* <http://www.nydailynews.com/new-york/isis-recruiting-lone-wolfterrorists-targettimes-square-bratton-article-1.1941687> ("[Daesh] the terror group responsible for the videotaped executions of two American journalists and a British aid worker are calling for "lone wolf" attacks on Times Square."); Peter Beinart, "What Does Obama Really Mean by 'Violent Extremism'?", (Feb 20, 2015) online: *Atlantic* <http://www.theatlantic.com/international/archive/2015/02/obamaviolent-extremism-radical-islam/385700/> ("terrorism, […] is available to people of all ideological stripes and which grows more dangerous as technology empowers individuals or groups to kill far more people far more quickly than they could have in ages past."); Kathy Gilsinan, "Is ISIS's Social-Media Power Exaggerated?", (Feb 23, 2015) online: *Atlantic* <http://www.theatlantic.com/international/archive/2015/02/is-isisssocial-media-power-exaggerated/385726/> ("The high-quality videos, the online magazines, the use of social media, terrorist Twitter accounts—it's all designed to target today's young people online, in cyberspace."); Michael Schmidt, "Canadian Killed in Syria Lives on as Pitchman for Jihadis", (July 16, 2014), online: *New York Times* <http://www.nytimes.com/2014/07/16/world/middleeast/isisuses-andrepoulin-a-canadian-convert-to-islam-in-recruitment-video.html?_r=0>; Ben Hubbard, "Jihadists and Supporters Take to Social Media to Praise Attack on Charlie Hebdo", *New York Times* (Jan. 11, 2015), online: <http://www.nytimes.com/2015/01/11/world/europe/islamicextremists-take-to-social-media-to-praise-charlie-hebdoattack> [http://perma.cc/CH92-LH5S] ("Within hours of the deadly attack on the French satirical newspaper Charlie Hebdo, Islamic extremists and their supporters were praising the killings and lauding the attackers on social media").

[80] JM Berger & Jonathon Morgan, "The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter" (2015) online (pdf): < https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf> ("During the period of October 4 through November 27, 2014, we estimate there were no fewer than 46,000 Twitter accounts supporting ISIS. (…) We estimate that a minimum of 30,000 of these are accurately described as accounts belonging to ISIS supporters and controlled by a human user, using the most conservative criteria.").

multiple platforms.[81] IS's 'retweet army' used tactics such as hijacking hot hashtags on subjects ranging from British soccer to California earthquakes to grab attention and distribute its inflammatory message.[82] As a result of these figures, government authorities in Western nations have issued warnings about social media-fueled terrorist attacks.[83] Furthermore, as previously mentioned cases of Myanmar and Sudan have demonstrated,[84] using social media's capacity to incite hatred has become a *modus operandi* amongst modern genocidaires. However, human rights jurisprudence and research on speech[85] appear to lack the dynamics and particularities of digital incitement, a possibility clearly unimagined by the twentieth century's drafters, prosecutors, and observers.

As previously stated, the limitations on freedom of speech is a very limited section of IHRL that mainly developed in response to state-sponsored propaganda efforts of the twentieth century. Private non-state actors with the ability to conduct effective hate campaigns outside of the government machinery were, in fact, completely unanticipated by the drafters of the leading human rights agreements.[86] This state-centric norm seeks to compromise the necessity for open and informed debate in a democratic society (individual autonomy) and the avoidance of discrimination and assaults on protected vulnerable groups.

---

[81] Eric Schmitt, "U.S. Intensifies Effort to Blunt ISIS' Messages", *New York Times* (Feb 16, 2015), online: <http://www.nytimes.com/2015/02/17/world/middleeast/us-intensifieseffort-to-blunt-isis-message.html> ("With the Islamic State and its supporters producing as many as 90,000 tweets and other social media responses every day").

[82] Uri Friedman, "An American in ISIS's Retweet Army", *Atlantic* (Aug. 29, 2014), online: <http://www.theatlantic.com/international/archive/2014/08/anamerican-in-isis-retweet-army/379208/>

[83] Scott Neuman, "Homeland Security Chief: Threat to U.S. Malls 'A New Phase' For Terrorists", *NPR* (Feb. 22, 2015), online: <http://www.npr.org/blogs/thetwoway/2015/02/22/388242488/homeland-security-chiefthreat-to-u-s-mallsa-new-phase-for-terrorists>.

[84] See discussion *above* in part 2.2 nature and impact of expressions on the Internet.

[85] Though a few academics are addressing this topic, most researchers do not consider digital platforms to be distinct from the rest of conventional media platforms.

[86] See Gordon *supra* note 49, O'Donnell *supra* note 78.

Moreover, the term "hate speech" is an over-used phrase still unclear under international law. On the one hand, its extent and lack of agreement on its definition make it vulnerable to state misuse and undue restriction of legitimate expression. It allows states to confuse the concept with "fake news," weaponizing international norms to suppress dissenters, activists, and political opponents.

Several treaties and declarations oblige states to prevent and prohibit 'hate speech in all media.[87] For example, article 7 of the UDHR prohibits and protects from any incitement of discrimination.[88] The ICCPR in its article 20(2) obliges states to prohibit by law "any advocacy of … hatred that constitutes incitement to discrimination, hostility or violence."[89] This obligation to "prohibit by law" does not necessarily mean criminalization.[90] The human rights committee has observed that this duty requires states to "provide appropriate sanctions" such as civil or administrative penalties.[91] To comply with article 20(2), a law must make the proper definition of propaganda and advocacy of hatred and the other related terms, and provide appropriate sanctions in case of any infringement.[92]

[87] As an aside, it is worth noting that there is a recurring structural conflict in the major human rights instruments between the prerogative of free expression and the prerogative of freedom from invidious discrimination. See Gregory S. Gordon, A War of Media, Words, Newspapers and Radio Stations': The ICTR Media Trial Verdict and a New Chapter in the International Law of Hate Speech (2004) 45 VA J Intl L 139 at 145–153 (analyzing the tension between and among the provisions of these international instruments). The Human Rights Committee, cognizant of this inbuilt clash, has even felt it necessary to stress that has sought to stress that Article 20 of the ICCPR (prohibiting hate speech) is fully compatible with the right to freedom of expression. *General Comment no. 11, Article 20, Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred*, Human Rights Committee (HRC), 19th Session, CCPR/C/GC/11, 29 July 1983 at para 2; *General comment no. 34, Article 19, Freedoms of opinion and expression*, UN Human Rights Committee (HRC), 102nd Session CCPR/C/GC/34, 12 September 2011 paras 48-52 (stressing that the provisions complement each other and that Article 20 "may be considered as *lex specialis* with regard to Article 19"). However, this structural tension is of limited relevance for this thesis as online hate speech largely raises the same free speech concerns as to its offline variant.

[88] *Universal Declaration of Human Rights*, 10 December 1948, G.A. Res. 217 A(III), U.N. Doc. A/810. art 7 [UDHR]

[89] ICCPR, *supra* note 1 art 20(2).

[90] General Comment no 11, *supra* note 87 at para 2.

[91] *Ibid.*

[92] *Ibid.*

One thing is pertinent to note here is that, the limitations provided in article 19(3) of ICCPR that, expressions may be restricted in … the respect of "the rights of others" is different from the previously mentioned article.[93] Thus ICCPR acknowledges the gravity of hate speech, and there is also the scope of assessment of the nature of the speech.[94] While it is quite clear on certain restrictions, it also provides room for situational assessment or ambiguity. The Human Rights Committee concludes that Articles 19 and 20 of the ICCPR "are consistent with and complement one another."[95] Regardless, they are still open to interpretation. The 2012 Rabat Plan of Action[96] provides some direction, advancing a series of authorities following consultation sessions conducted by the Office of the High Commissioner for Human Rights (OHCHR). It offers a six-part threshold assessment to identify grave hate speech that triggers states' obligations under Article 20(2), taking into account context, speaker, intent, substance, the intensity of the speech, and the likelihood of harm arising as a result of the speech.[97] Key terms are further defined as follows:

---

[93] Hate speech is more likely to violate equality rights, such as the right to be free from discrimination and the right to human dignity. See, Robert C. Post, "Racist Speech, Democracy, and the First Amendment" (1991) 32 William and Mary L Rev 267 at 272; Kevin Boyle, "Hate Speech— the United States Versus the Rest of the World?" (2001) 53 Maine L Rev 487 at 490.

[94] The Human Rights Committee, while deciding cases concerning Article 20 even avoided defining incitement of hatred. Human Rights Council, Implementation of General Assembly Resolution 60/25A of 15 March 2006 Human Rights Council: *Incitement to Racial and Religious Hatred and the Promotion of Tolerance: Report of the High Commissioner for Human Rights*, 2nd Sess, UN Doc. A/HRC/2/6, 20 September 2006 at para 36; Human Rights Council: Report of the United Nations High Commissioner for Human Rights and Follow-Up to the World Conference on Human rights, Addendum, Expert Seminar on the links between Article 19 and 20 of the International Covenant on Civil and Political Rights: "Freedom of expression and advocacy of religious hatred that constitutes incitement to discrimination, hostility or violence", 10th Sess, UN Doc. A/HRC/10/31/Add.3, 16 January 2009 at para. 1.

[95] General comment No. 34 (2011), *supra* note 59 at para 50.

[96] The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. Human Rights Council, Annual report of the United Nations High Commissioner for Human Rights Addendum Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, 22nd Sess, UN Doc. A/HRC/22/17/Add.4, Appendix, adopted 5 October 2012.

[97] A wide spectrum of UN Human Rights Council special procedures has adopted the Rabat Plan of Action. see, e.g. GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and, 7th Sess, UN Doc. A/67/357, 7 September 2012; GA, Report of the Special Rapporteur on freedom of religion or belief, 25th Sess, UN Doc. A/HRC/25/58, 26 December 2013; Report of the Special Rapporteur on contemporary

> "Hatred" and "hostility" refer to extreme and illogical emotions of opprobrium, enmity, and detestation towards the target group; "advocacy" is defined as the desire to publicly promote hatred towards the target group; and "incitement" refers to statements about national, racial, or religious groups that create an imminent risk of discrimination, hostility, or violence against members of those groups.[98]

It is worth noting that various international human rights norms give greater protection against discrimination than article 20 (2)'s emphasis on race and religion. For example, article 2(1) of the ICCPR ensures that all individuals' rights are protected. In contrast, article 26 expressly states that "the legislation should prohibit any discrimination and provide to all persons equal and effective protection against discrimination on any ground." International standards provide safeguards against discrimination based on race, color, sex, language, religion, political or other opinions, national or social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity, or intersex status.[99]

The spectrum of protection has widened over time. Other factors, such as age or albinism, are now explicitly protected. Given the global extension of protection, the rule against inciting should be interpreted to extend to the more significant categories presently covered by international human rights law. The ICERD contains a similar but more elaborate prohibition of hate speech. Article 4 incorporates the obligation to take "immediate and positive measures"

---

forms of racism, racial discrimination, xenophobia and related intolerance on manifestations of racism, racial discrimination, xenophobia and related intolerance, 26th Sess, UN Doc. A/HRC/26/49, 6 May 2014; and the contribution of the UN Special Advisor on the Prevention of Genocide to the expert seminar on ways to curb incitement to violence on ethnic, religious, or racial grounds in situations with imminent risk of atrocity crimes, 22 February 2013.

[98] See for reference, A/HRC/22/17/Add.4, *supra* note 96 appendix. Former UN Special Rapporteur Frank La Rue recognized whether there was a "real and impending threat of violence stemming from the speech" as a critical criterion in examining hate speech. (A/67/357, para. 46). See also Article 19, *Prohibiting Incitement to Discrimination, Hostility or Violence* (London, 2012), pp. 24–25.

[99] Article 19, "Hate Speech" Explained: A Toolkit (London, 2015), p. 14.

to eradicate hate speech and reinforce broader obligations under the Convention to devote the broadest possible range of resources to controlling and eliminating discrimination.[100] "Measures" are described by the CERD Committee as "legislative, executive, administrative, budgetary, and regulatory instruments...as well as plans, policies, programmes, and...regimes."[101] The scope of State responsibility under the ICERD is significantly greater than the ICCPR's obligation to "prohibit by law," and it may extend to regimes of Internet intermediary responsibility. Moreover, unlike the ICCPR's more severe term of 'advocacy of hatred,' which is interpreted to require the author's purpose to spread hatred,[102] the ICERD's ban encompasses any transmission of notions of racial superiority or hatred.[103] However, this substantial protection is confined to speech referring to race and ethnicity and excludes gender and sexual orientation.

The Committee for the Elimination of Racial Discrimination, the expert treaty-monitoring body for the ICERD, provides additional interpretive assistance. The Committee adopted the Rabat Plan of Action in 2013 by confirming that the "due respect" obligation under article 4 involves rigid compliance with freedom of speech safeguards.[104] It mainly states that the obligation of criminalization should be restricted to a small number of clearly defined and narrow instances:

[100] ICERD, *supra* note 56 at Art. 4 (directing States to "condemn all propaganda … based on ideas or theories of superiority of one race or group of persons of one color or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination …" as well as to criminalize "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another color or ethnic origin.").

[101] Official Records of the General Assembly, 64th Sess, Supplement No. 18 (A/64/18), annex VIII at para 13.

[102] UN Doc. A/HRC/2/6 *supra* note 94 at para 39.

[103] The CERD Committee has also established a set of similar factors to the Rabat Plan of Action for its suggestions on how to comply with Article 4's responsibility to restrict certain types of communication.

[104] Committee on the Elimination of Racial Discrimination, general recommendation No. 35 (2013), para. 19. According to the Committee, the due-regard clause is essential for freedom of expression. It describes it as "the most significant reference principle for calibrating the constitutionality of speech limitations."

> The criminalization of racist language should be kept for extreme situations that can be established beyond a reasonable doubt. At the same time, less severe cases should be dealt with through alternative ways, taking into account, among other things, the nature and intensity of the impact on targeted individuals and groups. Criminal punishments should be applied following legality, proportionality, and necessity criteria.[105]

The Committee has also limited the prohibition on "insults, mockery, or defamation of individuals or groups or justification of hate, contempt, or discrimination" to circumstances that "clearly amounts to incitement to hatred or discrimination."[106] This, particularly for social networks' speech, serves to confine the dangerously vast vocabulary of 'ridicule' and 'justification' to avoid restricting otherwise valid rights to mock and offend.

Although not exhaustive, this assessment argues that current human rights standards provide a suitable baseline against which local legislation and corporate actions can be measured. State parties and companies seeking advice under the ICCPR must, at the very least, enact effective legislation to condemn and limit severe forms of harmful online expression (through administrative or criminal consequences).

More generously, parties to the ICERD must take substantial steps and devote considerable resources–broadly defined–to abolish all kinds of racial hate speech. However, IHRL remains silent on numerous distinguishing aspects of online speech, leaving a governance vacuum that, as detailed below, is gradually being filled by non-state actors.[107] Is international law clear and strong enough to govern the unique characteristics of online hate speech? If not, in the absence

[105] *Ibid*, at para 12.
[106] *Ibid*, at para 13.
[107] Ido Kilovaty, "Are Tech Companies Becoming the Primary Legislators in International Cyberspace?" (28 March 2019) Lawfare, online:<https://www.lawfareblog.com/are-techcompaniesbecomingprimarylegislators-internationalcyberspace?fbclid=IwAR1T9o2T1KQn-RQWYgRY_pIyjXiByy1-Aw_aPrMkXrrf3Nz6uhbH15HhTxw#__prclt=pLTEhkPm>.

of such governing mechanisms, is IHRL providing the scope of governing cyberspace to the private entities?

2.6 Challenges with the Existing Mechanisms

When applying current international human rights norms to online communication, at least three obstacles and limitations emerge. Firstly, and perhaps the most significant one is about jurisdiction. States' human rights duties concerning hate speech are largely confined to those "within their jurisdiction."[108] The transnational nature of online platforms where any expression is expressed is creating the problem of definite territorial demarcation and posing cross-jurisdictional cooperation. The major human rights treaties and their corresponding jurisprudence are mostly quiet on the scope of State duties involving transnational internet communication. Whatever the case may be, state experience implies that territorial jurisdiction over online offences is fairly permissible.

The only international standard-setting document on the subject is the Council of Europe's Cybercrime Convention (2004), (mostly referred to as the Budapest Convention)[109] which states that the location of the "attacked computer" system is sufficient to establish the *locus delicti* in the issue.[110] Besides, the United Kingdom's Computer Misuse Act needs a significant relationship with domestic jurisdiction, such as the server's location at the crucial time.[111] The capitalist American approach establishes jurisdiction over offences committed against a "protected computer" that is "used in or affecting interstate or foreign commerce or

[108] See for ref. ICCPR, *supra* note 1 art. 2 (limiting a State's obligations to individuals "subject to its jurisdiction"); ICERD, art. 3, 6 for a comparable provision.

[109] *EC Convention on Cybercrime* (adopted 23 November 2001, entered into force 1 July 2004) 185 European Treaty Series.

[110] Explanatory Report of the Convention on Cybercrime (13 November 2001) CM (2001)144 addendum, para 233
<https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804d873c4

[111] UK Computer Misuse Act 1990, s 455

communication, including a computer located outside the United States that is used in a manner that affects interstate or foreign commerce or communication of the United States."[112] Singapore and Malaysia, on the other hand, only demand that internet data be located in their territory during the conduct of the crime,[113] or that the data be accessible in computers there.[114] On the one hand, all of this mainly establish the criminal law jurisdiction in cyberspace; on the other hand, it is not extraterritorial since its effects are being felt at home.

Secondly, as discussed further below, there are substantial concerns regarding the feasibility of effective compliance by private tech corporations with international human rights norms. Without going into an unduly pessimistic analysis, it is widely recognized that IHRL only imposes duties for States under traditional international law.[115] Article 20 of the ICCPR demands the legal ban of some types of hate speech, thereby directly addressing States. As a result, the willingness of states (particularly the United States, where most Information and Technology corporations are located) to enforce this body of norms will substantially impact the amount to which private actors will comply with such international standards. According to the OHCHR's Guiding Principles on Business and Human Rights, "the inability to implement existing laws that directly or indirectly govern business respect for human rights is frequently a serious legal vacuum in State practice."[116] The Guiding Principles on Business and Human Rights (also known as Ruggie principles) outline some obligations that corporations must fulfill to respect human rights in their operations. These include the need to

[112] The Computer Fraud and Abuse Act 1986,18 USC s 1030 (e) (2) (b)

[113] The Computer Misuse and Cybersecurity Act, Act 19 of 1993, s 11(3) (Revised 31 July 2007), as amended on 2 January 2011

[114] The Computer Crimes Act 1997, Act 563, as amended on 1 January 2006, s 9(2).

[115] A discussion was made regarding the platform governance by intermediaries (non-state actors) in section chapter 3. See *infra* 3.3.

[116] United Nations Human Rights Office of the High Commissioner, "Guiding Principles on Business and Human Rights" (2011), online (pdf): <https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf> 1 at p 5.

avoid contributing to adverse human rights impacts[117] and the obligation to do due diligence to detect possible human rights impacts of corporate actions.[118]

Moreover, this principle was developed keeping the mining companies in mind considering their various abusive human rights activities. While international human rights law does not bind companies directly, international criminal law was specifically designed to bind non-state actors.[119] However, the digital platforms voluntarily adopt human rights principles in their policies. Nonetheless, due to a lack of guidelines regarding the scope of respect from their side, most of their policies are not protecting human rights globally.[120]

Thirdly, due to its state-centric perspective, present IHRL is deafeningly quiet on the particular governance issues of developing IHRL-compliant liability regimes for internet intermediaries. This is since most unlawful hate speech instances are generated by third-party content providers or platform users, rather than the Intermediaries themselves.[121] This presents Pandora's box of legal and policy issues for which IHRL gives no direction. Perhaps notably, states have taken somewhat divergent ways of regulating this governance dilemma. Most importantly, Section 230 of the Communications Decency Act (CDA) (1996) in the United States provides internet intermediaries with blanket protection for information uploaded on their platforms.[122] In this respect, the US Court applied this rationale to prevent the application of the Supreme Court of Canada's (fairly progressive) worldwide injunction ruling in Equustek.[123] In conjunction with the First Amendment's worldwide unprecedented 'protection' and the general policy approach

[117] *Ibid, Principle 13.*
[118] *Ibid, Principle 17.*
[119] Emma Irving, "Suppressing Atrocity Speech on Social Media" (2019) 113 American Journal of International Law 256–261 at 257.
[120] See for details, chapter 3.3 *infra*
[121] Nicolas P Suzor et al, "Human rights by design: The responsibilities of social media platforms to address gender-based violence online" (2018) 11:1 Policy & Internet 83 at 84.
[122] Communication Decency Act, 47 USC 230 CDA.
[123] Alicia Loh, "Google v Equustek: United States Federal Court Declares Canadian Court Order Unenforceable" (16 November 2017), online: <https://jolt.law.harvard.edu/digest/google-v-equustek-united-statesfederalcourt-declares-canadian-court-order-unenforceable>.

of 'cyberliberalism,'[124] section 230 of the CDA effectively transforms the United States into a safe house for hate speech. This is in plain contrast to the European approach, in which the European Court of Human Rights (ECtHR) has ruled in a defamation dispute that Article 10 of the ECHR both allows and, at times, requires States to hold liable online digital platforms for defamatory user-generated content.[125]

2.7 Conclusion

The above discussions reveal that existing laws cannot answer all the concerns raised by digital platforms. As explained in this chapter, the main reason is that IHRL confers the obligation to govern speech mainly on states. The drafters of the core human rights instruments were unable to foresee the future. As a result, existing laws have difficulty resolving the question of governance of digital speech. Moreover, international human rights law confers no obligation (other than respect) on intermediaries even though they are one of the most influential key players in this system. Any solution is unlikely to govern digital space by applying human rights without conferring limited duties to intermediaries. In the next chapter, I turn to the role of intermediaries in this issue of governance of digital speech.

[124] Dominic McGoldrick & Therese O'Donnel, "Hate-speech law: consistency with national and international human rights law" (1998) 18:4 Leg Studies 453 at 455: "At one jurisprudential extreme stands the practice of the United States Supreme Court, which has come closest to the acceptance of racist speech as a price that has to be paid for maintaining the pre-eminence of status of freedom of expression." See also Yu Wenguang, "Internet Intermediaries' Liability for Online Illegal Hate Speech" (2018) 13:3 Frontiers of L in China 342 at p 344

[125] Columbia University, Global Freedom of Expression, "*Delfi v Estonia*", online: <https://globalfreedomofexpression.columbia.edu/cases/delfi-as-v-estonia/>. See *Delfi AS v Estonia*, No. 64569/09, (16 June 2015) ECHR at para 159: "…where third-party user comments are in the form of hate speech and direct threats to the physical integrity of individuals, the member States may be entitled to impose liability on Internet news portals if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties."

Chapter 3

Intermediaries and Platform Governance

3.1 Introduction

Intermediaries[126] are an integral part of speech governance in digital platforms. While promoting free expression globally, intermediaries also adopt several censorship mechanisms to protect their users from harmful content. This chapter deals with their platform governance, especially content moderation, to argue that intermediaries are struggling to apply human rights norms in their content moderation mechanism in the absence of any guiding principle. It is pertinent to mention that, intermediaries are primarily responsible for their content moderation and its outcome. However, the state's intervention in platform governance must not be overlooked when it comes to platform governance. This chapter addresses platform governance from both ends. It reveals no fixed set of rules to govern digital platforms in the absence of universally recognized guidelines for both intermediaries and the state regarding content moderation. As a result, it creates various challenges while governing expression on a global platform.

3.2 Content Moderation as a Tool to Regulate expression

The need to foster safer online environments is crucial, mainly when social media platforms are always at risk of becoming breeding grounds for harmful content. Social media networks practice content moderation to keep users secure from harmful content such as hate speech, violence, nudity, and online abuse. Content moderation, as previously defined, is "the organized practice of screening User-Generated content (UGC) posted to internet sites, social

[126] Christina Mulligan, "Technological Intermediaries and Freedom of the Press" (2013) 66 SMU L Rev 157; Felix T. Wu, "Collateral Censorship and the Limits of Intermediary Immunity" (2011) 87 Notre Dame L Rev 293; Danielle Keats Citron & Helen Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age" (2011) 91 BUL Rev 1435; Balkin, *supra* note 10; Balkin, *supra* note 4.

media, and other online outlets."[127] Content moderation operations check UGC to verify whether a certain piece of content, such as picture, video, technology, complies with the platform's regulations. This procedure may be carried out by both human moderators and Artificial Intelligence (AI)/ Machine Learning (ML)-enabled machine moderation. Trust and Safety (T&S) is an umbrella phrase for content moderation and other efforts to make online platforms safer.

To put it briefly, Content moderation is generally two types: *Ex ante* and *Ex post*.[128] In the former moderation, content is placed in a review queue before posting. The later type of moderation took place after posting content if any flagging occurred by other users or automatically.[129] Then there is a decision and procedure for appeal.

Online platforms employ an in-house staff of moderators and technology to moderate content, outsource the service to a service provider, or adopt a hybrid method with an optimal mix of both. Online content moderation (CM) outsourcing is a fast-increasing sector for various reasons. Increased demands for more moderation of the enormous volume of potentially harmful information being generated and shared online have provided new possibilities for providers to assist organizations in developing comprehensive CM and T&S policies and systems, sparking this industry.

3.2.1 Content Moderation by Intermediaries

Platform intermediaries are trying to ensure their respective platforms are safe for everyone. In order to do that, they are actively moderating content posted on their sites. Facebook, Twitter,

[127] Roberts, *supra* note 11.

[128] Klonick, *supra* note 12.

[129] Ralitsa Golemanova, "What Is Content Moderation? | Types of Moderation & Content to Moderate", (8 September 2021), online: *Imagga Blog* <https://imagga.com/blog/what-is-content-moderation/>.

and YouTube conduct content moderation from third-party outsourcing, while Tiktok directly employs content moderations using AI technology.[130]

All of these platforms[131] have their own rules of content moderation, and they are constantly upgrading to tackle new concerns such as extremist content, terrorism, cyberbullying, harassment, or revenge porn.[132] The traditional "public square" concept has been criticized for failing to safeguard vulnerable individuals who are pushed off sites by material such as hate speech and behaviour such as harassment. Digital platforms favour a "curated community approach," in which standards are framed as "we are a group, and we have a, b, and c ethical guidelines for treating one other."[133] This perspective was echoed by Facebook's Monika Bickert, who said in a report that, the business is not just attempting to "balance safety and free expression," but rather to set speech rules to "build a safe community."[134]

Platform entities are trapped between many competing pressures when it comes to standardization.[135] issues. On the one hand, they are being asked to publicize their comprehensive content moderation guidelines.[136] On the other hand, they are warned that public standards may be easily monetized, with offenders deliberately adjusting harassment to bypass moderation–a common problem, according to Anti-Defamation League spokesperson

[130] Katie Schoolov, "Why content moderation costs billions and is so tricky for Facebook, Twitter, YouTube and others", (27 February 2021), online: *CNBC* <https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html>.

[131] Several internet platforms are moderating content posted on their platforms. Above mentioned are to provide a context of the situation.

[132] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018) at 17.

[133] *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*, by Robyn Caplan (Melbourne: Data & Society Research Institute, 2018) at 12.

[134] Interview with Monika Bickert, head of global policy management from Facebook. *Ibid.*

[135] Standardization means platform transparency procedures should adhere to the same basic level of clarity when disclosing data. *See for more details:* Archon Fung, Mary Graham & David Weil, *Full disclosure: the perils and promise of transparency* (New York: Cambridge Univ. Press, 2007) at 59-63.

[136] Russell Brandom, "New rules challenge Google and Facebook to change the way they moderate users", (7 May 2018), online: *The Verge* <https://www.theverge.com/2018/5/7/17328764/santa-clara-principles-platform-moderation-ban-google-facebook-twitter>.

Brittan Heller.[137] In addition, they are increasingly being forced to make choices in response to removal demands from foreign governments.[138] This forces businesses to make a tough decision, which firms sometimes portray as a "delicate balance between respecting a foreign nation's sovereignty and submitting to government restrictions."[139] One of such platforms' representative remarked that the alternative option, intervening and making judgments that preserve ideals they may believe (such as defending the speech of LGBTQ users in regions where such speech is illegal), may be perceived as a Western entity projecting its ideology abroad.[140] The most significant issue observed in almost all the platforms discussed above is the difficulty in making policies to tackle hate speech and disinformation since they vary depending on local circumstances and power relations.

According to some mega-platforms, one of the most challenging areas of content moderation is hate speech because of the challenges involved in finding a definition of hate speech that can be applied internationally and at scale.[141] Companies struggle to manage geographical and cultural differences, dog whistles, and reclaimed words by disadvantaged minorities. Such issues are exacerbated by a desire to expand into regions of the world where they lack local moderator competence and may not even provide translations of their community rules into the local language.[142] Companies have created intricate methods for assessing hate speech based on defined protected features over time, but these systems also allow for a great deal of

[137] Interview with Brittan Heller from the Anti-Defamation League. Caplan, *supra* note 133 at 12.
[138] See for example: "Twitter receives record number of gov't requests to remove posts", *Al jazeera* (26 January 2022), online: <https://www.aljazeera.com/news/2022/1/26/twitter-sees-record-number-of-govt-demands-to-remove-content>.
[139] Caplan, *supra* note 133 at 12.
[140] Interview with Alex Feerst, head of legal at Medium. *Ibid*.
[141] Arcadiy Kantor, "Measuring Our Progress Combating Hate Speech", (19 November 2020), online: *Meta* <https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/>; Twitter Safety, "Updating our rules against hateful conduct", (13 December 2021), online: *Twitter Safety* <https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate>.
[142] Timothy McLaughlin, "How Facebook's Rise Fueled Chaos and Confusion in Myanmar" *Wired* (6 July 2018), online: <https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/>.

discretion in enforcement,[143] resulting in over-and under-removals of contents and suppressing free speech.[144] This limits the capacity of marginalized groups to use the platforms. Facebook's hate speech policy is the most complicated of the three platforms, involving a three-tiered system of prohibited content that distinguishes between protected and quasi-protected characteristics (i.e., race, ethnicity, national origin, religion or belief, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability) (i.e., age and immigration status).[145] The protected features lists on Twitter and YouTube heavily overlap with Facebook's.[146] Age is considered on par with the other protected factors in Twitter's list, and assaults based on immigration status are not protected.[147] The list on YouTube includes victims of a significant violent incident and veterans.[148]

Platforms do not adequately consider power dynamics in their rule formation when responding to hate speech, resulting in weird and irrational consequences.[149] An internal Facebook training

---

[143] There is a common way to define hate speech. See United Nations, "United Nations Strategy and Plan of Action on Hate Speech," (May 2019), online (pdf): *United Nations* <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOP-SIS.pdf.>

[144] Evelyn Douek, "More Content Moderation Is Not Always Better" *Wired* (2 June 2021), online: <https://www.wired.com/story/more-content-moderation-not-always-better/>.

[145] The first category includes violent or demeaning statements directed against a person based on a protected trait or immigration status. The second layer forbids assertions of inferiority directed at a specific individual because of a protected feature. The third layer includes requests for segregation, exclusion, or insults directed at a specific individual because of a protected feature. Facebook likewise safeguards against age-based assaults, but only when age is combined with other protected category. "Hate Speech | Transparency Center", (8 February 2022), online: *Meta* <https://transparency.fb.com/policies/community-standards/hate-speech/>.

[146] The Twitter rules states: "You may not promote violence against, threaten, or harass other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." "The Twitter rules: safety, privacy, authenticity, and more", (8 February 2022), online: *Twitter Help Center* <https://help.twitter.com/en/rules-and-policies/twitter-rules>; "Hate speech policy - YouTube Help", (8 February 2022), online: *YouTube Help* <https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436>.

[147] "Twitter's policy on hateful conduct | Twitter Help", (8 February 2022), online: *Twitter Help Center* <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>.

[148] The YouTube Team, "Our ongoing work to tackle hate", (5 June 2019), online: *blog.youtube* <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/>.

[149] See for example: Faiza Patel & Laura Hecht-Felella, "Facebook's Content Moderation Rules Are a Mess | Brennan Center for Justice", (22 February 2021), online: The Brennan Center for Justice <https://www.brennancenter.org/our-work/analysis-opinion/facebooks-content-moderation-rules-are-mess>.

document from 2017 indicated, for example, that only white males would be protected under the company's hate speech policy, out of three groups: female drivers, Black children, and white men.[150] The reasoning was that race (white) and gender (male) are protected traits. However, the other cases included quasi-or nonprotected qualities, such as age (in the case of Black children) and driving (in the female drivers' example).[151] When Facebook's hate speech policy was implemented, it resulted in the platform strengthening safeguards for white males, a dominant group, while failing to address speech targeting more vulnerable populations (women and Black people). Following the publication of the training papers, Facebook said that it has modified its hate speech enforcement mechanisms to deprioritize remarks regarding "Whites," "males," and "Americans."[152] However, it did not modify its core policies, and it could not demonstrate how it executed these revisions or how they were evaluated for success.

3.2.2 Content Moderation: Without Context?

The problem of content moderation by intermediaries is that it often results in the suppression of free speech.[153] The reason is a failure to take the "context" into their decision-making process. Context varies depending on the social, political, economic, and cultural conditions of a particular society. The context must be considered when moderating, for example, the user's location and when something was posted. Fact-checkers (moderators), artificial intelligence (AI), and other technological filters are responsible for taking down any content posted on an

[150] Julia Angwin & Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children", (28 June 2017), online: *ProPublica* <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms?token=sDXc9-_dthGWtDlWPbCrrxBIUqQRIggX>.
[151] *Ibid*.
[152] Adam Smith, "Facebook comments like 'white men are stupid' were algorithmically rated as bad as antisemitic or racist slurs", (4 December 2020), online: *The Independent* <https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-comments-algorithm-racism-b1766209.html>.
[153] Douek, *supra* note 144.

internet platform. However, AI and Machine Learning (ML) are not sufficiently capable of accessing the context of the content. Human intervention is needed to access critical content.

However, even human moderators are not always capable of accessing the context of the content. For example, a Cuban journalist was subjected to suppression of free speech. In a country like Cuba, where there is no respect for press freedom, his post relating to the corruption was taken down by Facebook without explanation.[154] In this kind of situation, internet platforms work or should work as a medium of last resort. For people like Cuban journalist, social media platform is a means to grab attention nationally and globally to show what is happening in their part of the world.

The need to strike a balance between rule consistency and sensitivity to local circumstances, particularly for concerns such as hate speech and misinformation, is relevant when addressing platform design and scale challenges.[155] Context preservation is a big problem on platforms that tend to collapse it at every step,[156] both in terms of how individuals acquire information (a post from a friend and a message from a news agency tend to seem quite similar) across cultures with varied histories and power dynamics, as well as in the receipt of information by both other users and moderators. Simultaneously, preserving consistency between decisions is required philosophically, for example, to ensure rules are not imposed arbitrarily or to provide some sense of "fair," and practically, when thousands of people are on-boarded to handle content problems. Maintaining this balance is not unique to platform entities; national and international

[154] Cindy Harper, "Facebook censorship hinders Cuban journalists reporting on corruption", (28 November 2021), online: *Reclaim The Net* <https://reclaimthenet.org/facebook-censorship-hinders-cuban-journalists-reporting-on-corruption/>; Rachel Bovard, "How many times must Facebook be caught censoring the truth?", *New York Post* (22 November 2021), online: <https://nypost.com/2021/11/22/how-many-times-must-facebook-be-caught-censoring-the-truth/>.

[155] Michael Herz and Peter Molnar, *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (Cambridge, UK: Cambridge University Press, 2012).

[156] Henry Jenkins, *Convergence Culture: Where Old and New Media Collide* (New York: New York University Press, 2006).

law have traditionally battled with adjusting restrictions on the material, such as hate speech, to different traditions and histories, while staying "law-like" and maintaining baseline standards.[157] Nonetheless, content policies are not laws; they are policies. This allows platform entities more latitude in enforcing their rules while also obscuring the policy formulation and enforcement process from public scrutiny.

This problem in content moderation extends beyond setting rules to enforcing them. Content is frequently reviewed outside of the environment in which it is created, especially when an organization grows in size. To effectively judge whether material is hateful, a moderator must understand the context in which it was created, including information about the creator, the target, and the setting, as well as language or cultural cues that they may not have access to (such as sarcasm, or newsworthiness). Moderators also must be highly self-aware of their environment. Moderators must not assume that viewers see a specific post contrasted against the same hate speech, pornography, and vulgar comedy that they just evaluated as they make their way through the moderation queue.[158]

Because of workload and work expectations, moderators need to respond to all of these criteria in a matter of seconds (or less) frequently. One former Facebook employee said that making content management more regional and responsive frequently included making judgments based on incomplete information. "Who is historically disadvantaged to whom is context-dependent and situational," he said, referring to the difficulty of analyzing hate speech directed against a Japanese person as an example. "Are you historically disadvantaged because of Japanese imperialism in China, or are you historically disadvantaged because of the treatment of Japanese Americans in the United States?" must be considered by moderators.[159] To solve

[157] Herz, *supra* note 155.

[158] Moderation queue means the list of posts to be checked by the moderators. In other words, the initial fate of content is waiting for approval. (Everyone expect their post to be approved!)

[159] Interview with Craig Colgan (pseudonym), a former employee at Facebook. Caplan, *supra* note 133 at 14.

these context problems at the time (a mere 70 million users compared to today's 4.62 billion), "you would have to employ everyone in India to look at all the published stuff, and you still would not be able to accomplish it."[160]

As a result, several platforms, including Twitter, are looking for new "signals" to monitor material (e.g., comments and interactions) that may be used to draw attention to problematic issues. However, these can also cause context challenges. According to one Twitter employee, likes and responses might indicate a variety of things in different contexts, making judgments challenging to automate:

> People seek attention in similar ways. A spammer seeking attention resembles a rapper attempting to release their newest mixtape, while the people respond like someone attempting to participate in a targeted harassment campaign. Just because something is well-coordinated does not make it unpleasant.[161]

The organization of the content moderation team has a considerable impact on how a platform manages these conflicts. There are some similarities in how moderators were notified of and dealt with inappropriate content across small, giant, and medium-sized teams. However, there are substantial divergences in how these teams can adjust to cultural variances, such as language gaps in the material, and their ability to use artificial intelligence to automate content flagging and removal.[162] When considering potential tools to monitor these corporations as they make critical decisions concerning the future of online speech, consideration must be

[160] *Ibid*.
[161] Interview with Del Harvey, vice president of Trust & Safety at Twitter. *Ibid*.
[162] *Ibid* at 23.

given to organizational dynamics and the tradeoffs that companies make, which are frequently concealed from public view.[163]

Global debate in content moderation has mainly focused on a small number of larger corporations – particularly Facebook (Meta), Twitter, and Google (mainly YouTube) – that have been labeled "industrial"[164] owing to their magnitude and quantity of users, the size of their content moderating staff, the operationalization of rules, and the separation of policy and enforcement at their organizations.

These larger corporations often started with the artisanal content management style and utilized this type of experimentation to build more structured, static, and inflexible rules. A portion of this formalization has happened due to fast expansion and the necessity to train people who are being on-boarded in large numbers. These employees frequently make content judgments outside of the context of the original post.[165] To ensure fair and consistent judgments, complicated philosophical principles about what constitutes harassment, hatred, or truth must frequently be broken down into smaller, more interpretable components. According to one of the Facebook employees, the objective for these firms is to establish a "decision factory" that looks more like a "Toyota factory than it does a courtroom in terms of actual moderation."[166] Complex concepts such as harassment and hate speech are operationalized to ensure that these concepts are applied consistently across the organization.[167] He described the method as "trying to take a complex process and break it down into really little components, so that you can

[163] Joan Donovan, "Why social media can't keep moderating content in the shadows", (6 November 2020), online: *MIT Technology Review* <https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/>.
[164] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (New Haven: Yale University Press, 2018).
[165] Caplan, *supra* note 133 at 24.
[166] Interview with a former employee of Facebook. *Ibid*.
[167] Gillespie, *supra* note 164.

routinize repeating it over and over and over again."[168] In this sense, industrial organizations are large-scale bureaucracies with highly specialized teams with duties and powers distributed. As Gillespie has pointed out, the spread of labour, which is frequently dispersed across the company and the globe, creates logistical challenges in information transmission about changing policies. This involves conveying information regarding the efficacy or correctness of policies to business policymakers.[169] Because of their magnitude, these corporations operationalize their content standards; the sheer volume of information that must be examined is challenging to comprehend. According to Nora Puckett, the YouTube representative at the 2018 Content Moderation at Scale Conference in Washington, D.C., YouTube removed 8.2 million videos from 28 million flagged in the fourth quarter of 2017, with 6.5 million flagged by automated means, 1.1 million flagged by trusted users, and 400,000 flagged by regular users. According to the same official, YouTube's content moderation teams employ 10,000 people. Despite being overshadowed by behemoths Facebook and Google, Twitter continues to have more than 330 million monthly users and billions of tweets every week. During the Content Moderation at Scale event, Del Harvey, vice president of Trust & Safety at Twitter, remarked that detecting 99.9 percent of harmful content still implies that tens of thousands of problematic tweets persist.[170] Similarly, Facebook's moderators make 300,000 mistakes daily.[171]

To highlight content such as hate speech, industrial content moderation teams are increasingly relying on automated techniques. Both Facebook and YouTube have acknowledged that they are now utilizing algorithms to locate objectionable content and remove it using "detection technology" before people report it, even if this content is still subject to human review.[172] In

[168] *Ibid.*
[169] *Ibid.*
[170] Del Harvey, Content Moderation at Scale Conference, Washington D.C. on May 7th, 2018.
[171] John Koetsier, "Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day", (9 June 2020), online: *Forbes* <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/>.
[172] note 145.

previous revelations about automated content takedowns, Facebook revealed that they had recorded high rates of success for this detection technology in the categories of graphic violence (86%), nudity, adult material (96%), and spam (100%).[173] The percentage of success of automated technology in detecting hate speech is lower, but still considerable, with detection technologies detecting and flagging "approximately 38% of the content they took action on for hate speech, using automated means."[174] As a result, rates of automated takedown are significantly higher for content types that the company considers less ethically ambiguous, such as spam/malware, child pornography, and terrorist propaganda (which requires its investigation into how companies categorize this type of content and the extent of false positives). Companies may be researching the use of automated technology in these other fields, as indicated by Facebook's identification of concerns such as hate speech.

3.3 The Struggle to Find the Balance: Diverse Culture and Liability in the Context

One of the significant challenges from the preceding discussion is that intermediaries face challenges in content moderation, mainly regarding the context. When global platforms reach the size of Facebook, Twitter, or YouTube, preserving consistency in decision-making frequently comes at the price of being localized or contextual. When making a moderating choice based on specific cultural and political circumstances, this might lead to issues with material such as hate speech, discrimination, or misinformation. Perhaps, as a result, platforms of this magnitude tend to collapse contexts in favour of developing universal norms that make little sense when applied to material from radically varied cultural and political settings throughout the world.

[173] Caplan, *supra* note 133 at 24.

[174] "Community Standards Enforcement | Transparency Center", (8 February 2022), online: *Meta* <https://transparency.fb.com/data/community-standards-enforcement/>; Caplan, *supra* note 133 at 24.

This can have a severe detrimental influence on marginalized populations at times. When Facebook sought to establish a policy that embraced conceptions of intersectionality abstracted from current power relations, basically defending the hegemonic groups of White and males but not "Black children," Julia Angwin attacked this sort of policy practice.[175] Her research indicated that attempts at universal anti-discrimination laws frequently fail to account for power disparities along racial and gender lines. In other cases, the Venus of Willendorf may be prohibited inadvertently for being too "pornographic."[176] The challenge of an under-resourced moderation system working under immense pressure is to apply standard linguistic criteria across nations, 111 official languages, and thousands of dialects.[177] The local context is often overlooked in this sort of moderation. The anti-Hezbollah demonstration in Lebanon a few years ago is a classic example of such a type of moderation. Videos posted to YouTube showed protesters yelling in Arabic; a moderator recognized the phrase Hezbollah but not much else in the language, and categorized the clip as content advocating this outlawed group.[178] It was removed, and the protesters' voices were suppressed. Failure to handle context concerns can have catastrophic repercussions. Tragically, this has been witnessed in the violence that has erupted in Myanmar, which has undoubtedly been driven by disinformation and hate speech propagated on both the Facebook network and its messaging service WhatsApp.[179] Facebook CEO Mark Zuckerberg admitted in April 2018 that the firm lacks the language and cultural

[175] Angwin & Grassegger, *supra* note 150.
[176] Aimee Dawson, "Facebook censors 30,000 year-old Venus of Willendorf as 'pornographic'", (27 February 2018), online: *The Art Newspaper* <https://www.theartnewspaper.com/2018/02/27/facebook-censors-30000-year-old-venus-of-willendorf-as-pornographic>.
[177] Kate Maltby, "The Online Harms Bill is a threat to freedom of expression, but it's buried in Government chaos", (28 January 2022), online: *inews.co.uk* <https://inews.co.uk/opinion/online-harms-bill-threat-freedom-expression-buried-government-chaos-1427925>.
[178] *Ibid.*
[179] Anthony Kuhn, "Activists In Myanmar Say Facebook Needs To Do More To Quell Hate Speech", *NPR* (14 June 2018), online: <https://www.npr.org/2018/06/14/619488792/activists-in-myanmar-say-facebook-needs-to-do-more-to-quell-hate-speech>.

tools to combat hate speech in the region.[180] Reuters reports that hate speech directed against the Rohingya population is still widespread throughout Facebook-owned and managed platforms despite his commitment to recruiting additional Burmese speakers.[181]

The struggle to consider cultural contexts makes content moderation even more complex for intermediaries. As argued earlier, intermediaries develop their policies to cope with emerging problems, and most of them are unknown to the public. Their policies cannot be termed as "law"; however, those policies are no less powerful than any law. Their policies determine which content stays online and which does not. Their policy determines who gets to speak and who does not. Additionally, in the absence of any universally accepted guiding principles for this type of media, intermediaries are making their policies, and those policies vary from one platform to another.[182] It is intermediaries in close connection to states that are suppressing freedom of expression. Because of their advertisement-driven (Ad-driven), business model intermediaries are forced to comply with states' policies and laws even if those policies result in the suppression of free speech.[183]

Not only are those laws imposing threats to freedom of expression, but they also impose certain liabilities and active intervention in the content moderation process by the intermediaries. The

[180] Steve Stecklow, "Why Facebook is losing the war on hate speech in Myanmar", *Reuters* (15 August 2018), online: <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.
[181] *Ibid.*
[182] The Santa Clara Principles on Transparency and Accountability in Content Moderation is a principle for ensuring transparency and accountability in content moderation is developed by civil society organizations. This cannot be termed universally accepted principles because of the non-involvement by states.
[183] Redaction AfricaNews, "Rwanda silencing YouTubers with 'abusive' legal framework - Human Rights Watch", (16 March 2022), online: *Africanews* <https://www.africanews.com/2022/03/16/rwanda-silencing-youtubers-with-abusive-legal-framework-human-rights-watch/>; Nwachukwu Egbunike & Kofi Yeboah, "Twitter's deal with Nigerian government sacrifices digital rights", (17 January 2022), online: *Global Voices* <https://globalvoices.org/2022/01/17/twitters-deal-with-nigerian-government-sacrifices-digital-rights/>;
Tomuwa Ilori, "In Nigeria, the government weaponises the law against online expression", (17 December 2021), online: *Global Voices* <https://globalvoices.org/2021/12/17/in-nigeria-the-government-weaponises-the-law-against-online-expression/>; Maltby, *supra* note 177; Derek de Preez, "UK's Online Safety Bill - not robust enough to to tackle illegal content, nor does it protect freedom of expression", (24 January 2022), online: *Diginomica* <https://diginomica.com/uks-online-safety-bill-not-robust-enough-tackle-illegal-content-nor-does-it-protect-freedom>.

European approach regarding platform governance is worth mentioning in this regard. The EU directive establishes intermediaries' role not necessarily as an editor, but allows them to be liable in case of any infringement because of their relationship with the content. An intermediary may not be held liable if there is no active involvement in the public transmission of unlawful content or if it is not aware of the infringing nature of the content. However, they must remove such content after being aware of such infringement.[184] According to Art. 15, intermediaries may not be subject to a general monitoring obligation to identify illegal activities. While interpreting this Directive in *Google France v. Louis Vuitton*[185] the court held that storage providers are exempted from liability because of not having "an active role of such a kind as to give it knowledge of, or control over, the data stored."[186] This Directive was further interpreted in *L'Oréal SA and others v. eBay International and others*[187] where the court held that the operator of a website is not liable for any content uploaded by a client because its role in such a case is neutral. However, if it played an active role in this process, it cannot claim such exemption as mentioned in Art 14(1) of the Directive 2000/31.[188] The Directive makes a clear distinction between the digital platforms and traditional media platforms (which is described as "on-demand media") regarding their liability and their relation with the content published on their respective platforms. Unlike the digital service providers and platforms, the providers of on-demand media services bear editorial responsibility for the content they publish, order and purchase since they have the final say in publishing a piece of content.[189]

[184] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the internal market ('Directive on electronic commerce', 'E-Commerce Directive'), 2000 O.J. (L 178), arts. 12–14.
[185] *Google France S.A.R.L. and Google, Inc. v. Louis Vuitton Malletier S.A. and Others,* Joined Cases of C-236/08 to C-238/08, ECLI:EU:C:2010:159. Cour de Cassation [Final court of appeals] *Google France S.A.R.L. and Google, Inc. v. Louis Vuitton Malletier S.A. and Others,* Joined Cases of C-236/08 to C-238/08, 23 March 2010.
[186] *Ibid,* para 120.
[187] *L'Oréal S.A. and Others v. eBay International A.G. and Others*, Case C-324/09, ECLI:EU:C:2011:474.
[188] *Ibid, para 116.*
[189] Directive, *supra* note 184 art. 1

There was no scope to include social media platforms in the AVMS directive. However, a recent amendment (adopted in 2018) includes audiovisual content published on social media platforms.[190] The new provisions of the Directive appear to be detailed, and the major platform providers have already taken steps to comply with those requirements that have now become mandatory.[191] The regulation only applies to a narrow range of content—specifically, audiovisual content—and the government is only granted control over platform providers' operations in connection with a handful of content-related issues, such as minor protection, hate speech, support for terrorism, child pornography, and denial of genocide.[192] In any event, such content is frequently blocked or removed by platforms upon obtaining knowledge of it in accordance with their regulations. Nonetheless, not all forbidden content in Europe is incompatible with such regulations. Platform providers will be compelled to take action under the E-Commerce Directive and the AVMS Directive once the Directive's provisions are implemented into national law in the EU Member States. These two pieces of law largely operate in conjunction, since the former mandates illegal content to be deleted in general, while the latter identifies specific categories of the infringing content and lays out comprehensive regulations for their removal. The AVMS Directive has several measures that ease the application of the requirements and serve as procedural protections.

Germany passed the Netzwerkdurchsetzungsgesetz (NetzDG), which imposes the same limits on hate speech online previously imposed on traditional media.[193] The rule, enacted in 2017, applies to "profitable" social media sites and "platforms producing journalistic or editorial

[190] The recent amendment introduced "video-sharing platform service" and "video-sharing platform provider". According to the amendment, social media platforms will fall under the scope of these terms despite their somewhat misleading name/terms.

[191] András Koltay, "The Protection of Freedom of Expression from Social Media Platforms" (2022) 73:2 Mercer L Rev 523–589 at 540.

[192] Koltay, *supra* note 191.

[193] Claudia Haupt, "Online Speech Regulation: A Comparative Perspective," Presented at the American Political Science Association, August (2018).

material" with over 2 million registered users who receive more than 100 complaints about "unlawful content" every calendar year.[194] The legislation imposes reporting requirements on the treatment of unlawful content, clear methods for handling complaints, and auditing rules mainly aimed at the organization of content moderation and trust and safety teams. The law clarifies that a platform must have 2 million users to be subject to these restrictions and provides exemptions for sites with less than 2 million users and nonprofit-making corporations.[195]

US law distinguishes platforms not by size, but rather by the designations "interactive computer services" and "publisher."[196] Due to Section 230 of the Communications Decency Act, platforms ("interactive computer services") are immune from liability for most sorts of non-illegal, non-copyrighted information. Furthermore, ISPs are permitted to freely "restrict access to or availability of content that the provider or user finds to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[197] This rule allows platforms to create and enforce their community norms as they deem fit. Proponents of the legislation argue that the tech sector as we know it would cease to exist if this clause did not exist. According to Eric Goldman, it is a "globally unique approach" that has given the United States a competitive advantage on the internet.[198] According to Jack Balkin, this provision is "among the essential free expression rights in the United States in the digital age."[199] Critics of the bill argue that the liability shield

[194] *Gesetz zur Verbesserung der Rechtsdurchsetzung in den sozialen Netzwerken* [Act to Improve Enforcement of Law in the Social Networks], BGBl. I, S. 3352 of Sept. 1, 2017, (Netzwerkdurchsetzungsgesetz, "NetzDG"). English translation by the Federal Ministry of Justice available at https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=92275CBB36905E837DBADFEEE79A0533.1_cid324?__blob=publicationFile&v=2

[195] It is not clear if this refers to 2 million German users or users globally. Many of the social platforms mentioned this uncertainty as to why they would not be subject to the law, while others were unsure if it would apply.

[196] *The Communications Decency Act* (CDA) of 1996, 47 U.S.C. Section 230 (c)(2)(A).

[197] Caplan, *supra* note 133 at 27.

[198] Eric Goldman, "The Ten Most Important Section 230 Rulings" (2017) 20 Tul J Tech & Intell Prop 1.

[199] Balkin, *supra* note 10.

for platforms is very broad,[200] and that additional restrictions and controls are needed to limit online defamation.[201]

One of the key reasons for the EU and US's opposing approaches is the level of free speech protections across the Atlantic. Serious threats to fundamental rights in Europe might be viewed as an example of states' affirmative need to regulate private activity in order to defend fundamental rights, as emphasized by the European Court of Human Rights.[202] On the question of liability of the intermediaries, the EU has made a progress and even the US is rethinking about their approach.[203] Moreover, the liability of intermediaries is yet to be established in international human rights law. The European approach might be the one that can guide in establishing a general liability and guiding framework of intermediaries globally.

Intermediaries are adopting various directives and regulations imposed by states and governments. Facebook, one of the intermediaries, has recently introduced an advisory body in their content moderation.

3.3.1 Facebooks' Oversight Board: An Attempt to Adapt to Cultural Diversity?

As a transnational or rather global platform, Facebook moderates content uploaded from almost all over the world. While moderating, many contents (sometimes even users) get removed/blocked by the system (AI, human moderators) only because of a failure to understand the context of such content. As a result, users' voice gets suppressed. Alternatively, the opposite problem exists. Content must be removed is not removed. In order to bring some transparency

[200] Kathleen Ann Ruane, "How Broad A Shield? A Brief Overview of Section 230 of the Communications Decency Act", (2018), online: *Congressional Research Service*.

[201] Ann Bartow, "Section 230 Keeps Platforms for Defamation and Threats Highly Profitable", (10 November 2017), online: *The Recorder* <https://www.law.com/therecorder/sites/therecorder/2017/11/10/section-230-keeps-platforms-for-defamation-and-threats-highly-profitable/>.

[202] See for example, *Von Hannover v Germany*, [2005] 40 EHRR 1; *Verein gegen Tierfabriken Schweiz (VgT) v Switzerland* [2001] 334 EHRR 159.

[203] "DEPARTMENT OF JUSTICE'S REVIEW OF SECTION 230 OF THE COMMUNICATIONS DECENCY ACT OF 1996", (3 June 2020), online: <https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996>.

to its content moderation system, Facebook has recently introduced an "oversight board" in its content moderation system.

The Board is established to promote free expression by making principled, independent decisions on the material on Facebook and Instagram and offering recommendations on the appropriate Facebooks' content policy.[204] Currently, the Board consists of 20 members, and the Board will have 40 members from all around the world representing a wide range of specialties, cultures, and backgrounds.[205] These members will be able to choose which content cases to evaluate and whether to sustain or reverse Facebook's content rulings. The Board is not intended to be merely an extension of Facebook's existing content approval procedure. Instead, it examines a small number of apparent instances to see if choices were made in conformity with Facebook's declared principles and standards.[206]

Facebook is a global platform promoting free expression globally.[207] Moreover, at the same time, its establishment of an oversight board is an example of the challenges of respecting cultural diversity and translating freedom of expression locally.

The Oversight Board rejected Facebook's decision to delete a Burmese post under its Hate Speech Community Standard.[208] The Board concluded that the post did not target Chinese individuals, but rather the Chinese government. It utilized obscenity to describe Chinese government policies in Hong Kong as part of a political debate on the Chinese government's role in Myanmar.[209] The background of this decision is: a Facebook user who appeared to be in Myanmar posted in Burmese on their timeline in April 2021. Following the coup in

[204] "Oversight Board | Independent Judgment. Transparency. Legitimacy.", online: <https://oversightboard.com/>.
[205] "Meet the Board | Oversight Board", online: <https://oversightboard.com/meet-the-board/>; note 204.
[206] note 204.
[207] Their design of platforms, features, auto-translation to reach a global audience, subtitles etc., are examples of the promotion of free expression worldwide.
[208] Facebook Oversight Board: Case decision 2021-007-FB-UA 11 August 2021
[209] *Ibid.*

Myanmar on February 1, 2021, the post examined measures to reduce funding to the Myanmar military. It advocated donating tax income to the Committee Representing Pyidaungsu Hlutaw (CRPH), a group of parliamentarians who opposed the coup. The post had about 500,000 views, yet no Facebook users reported it. The alleged infringing portion of the user's post was translated by Facebook as "Hong Kong people, because the f**ing Chinese tortured them, changed their banking to UK, and now (the Chinese) they cannot touch them."[210] Facebook deleted the post in accordance with its Hate Speech Community Standard.

The Board recommends Facebook "to ensure that its Internal Implementation Standards are available in the language in which content moderators review content. If necessary to prioritize, Facebook should focus first on contexts where the risks to human rights are more severe."[211]

In another case, The Board overturned Facebook's decision to delete a video of Colombian demonstrators condemning the country's president, Ivan Duque.[212] The demonstrators in the video use a term that is classified as a slur under Facebook's Hate Speech Community Standard. The Board recommended, "to publish illustrative examples from the list of slurs designated as violating under its Hate Speech Community Standard, including borderline cases with words which may be harmful in some contexts but not others."[213]

In both cases mentioned above, Facebook lacks the contextual understanding of the "terms" and lacks local understanding of the situation. They solely rely on their translation. Another example of not understanding the context is a post from a Russian user. On appeal to the Board, it rejected Facebook's decision to delete a comment in which a follower of imprisoned Russian opposition leader Alexei Navalny referred to another user as a "cowardly bot."[214] Facebook

[210] *Ibid.*
[211] See for more details: *Ibid.*
[212] Facebook Oversight Board: Case decision 2021-010-FB-UA 27 September 2021.
[213] See for more details: *Ibid.*
[214] Facebook Oversight Board: Case decision 2021-004-FB-UA 26 May 2021.

banned the remark because it included the term "cowardly," which was seen as a negative character accusation. While the elimination was under the Bullying and Harassment Community Standard, the Board determined that the present Standard was an unnecessary and excessive limitation on free expression under international human rights norms. It was also contrary to Facebook's ideals.[215]

The oversight board has been functioning since January 2021. Originally the idea was that the oversight board would decide Facebook's actual decision relating to its content moderation if it were in accordance with Facebook standards.[216] This is simply an internal exercise. However, the Board has also used international human rights as a guiding principle in its decision-making, which makes sense given Facebook's commitment to the UNGP.[217] So far, the majority of the judgments have dealt with hate speech in its broadest meaning, including what fits within Facebook's community standards for "Dangerous Individuals and Organizations."[218] The Board has criticized Facebook's guidelines for being overly broad and ambiguous in various instances.[219] This is particularly relevant to the "Community Standards on Dangerous Individuals and Organizations," and the Board generally criticizes the somewhat haphazard communication with various standards in various places, internal standards not communicated to the public, continuous alterations of the standards, and lack of translation into the appropriate language.[220] This indicates the problematic policy adopted by Facebook.[221]

The Board is still in its early days. However, most of its decisions so far are human rights friendly. Moreover, the diverse background of the members of the Board is an indication of

[215] See for more details: *Ibid.*
[216] Oversight Board Charter, Section 2. Basis of Decision Making.
[217] Schaumburg-Müller, *supra* note 12 at 21.
[218] Facebook Oversight Board: Case Decision 2020-001-FB-UA 28 January 2021, 2020-02, 2020-03, 2020-05 (all 28 January), 2020-07 12 February 2021-02 13 April, 2021-03 29 April.
[219] Facebook Oversight Board: Case Decision 2021-1, 2021-3, 2020-6, 2020-5, 2020-4.
[220] Facebook Oversight Board: Case Decision 2021-3 ('RSS is the new threat').
[221] Schaumburg-Müller, *supra* note 12 at 22.

accommodating representatives from various cultures taking decisions regarding content moderation on a global platform.

3.4 Conclusion

This chapter reveals the problematic content moderation by the intermediaries. In the absence of any internationally recognized normative framework for content moderation, they are creating their policies, and in most cases, they are different from one another. At the same time, most of their policies regarding content moderation remain in the darkness. Another problem with their content moderation is not being able to consider contextual aspects of content. Moreover, there are no universally accepted guidelines to regulate digital platforms. As a result, states are forcing intermediaries to comply with their own rules and regulations. While struggling to comply with various rules and regulations, they tend to remove more content than usual, which results in the suppression of freedom of expression. To overcome this, Facebook has introduced an Oversight Board to include diverse perspectives and expertise in its content moderation system. The next chapter sums up the whole thesis with possible solutions to this problem.

Chapter 4

Concluding Chapter

Final thoughts and summing up

This thesis discusses the nature and impact of expressions expressed on internet-based digital platforms. The discussion reveals that digital platforms are different from our traditional understanding of media in many ways and that their impact is far more reaching than other traditional media. Initially, I sought to draw an analogy on how the media has played a crucial role in developing international law, especially in the post-WW2 era. However, digital platforms are a very different kind of media. The fact that they are still governed by laws drafted in a period when the internet did not exist is a problem. This thesis has addressed some of the shortcomings of existing international laws governing freedom of expression in digital platforms with several examples in recent times.

The third chapter discussed the role of key players in this context. Intermediaries are the most crucial actors in this regard. They are *de facto* primarily responsible for the freedom of expression of their respective platform's users. They govern their respective platforms and decide which content stays or gets removed from their respective platforms. The system by which they make such decisions is commonly known as content moderation. In the absence of any internationally accepted guidelines, intermediaries make their content moderation policies, and most of such policies are invisible to the public eye. At the same time, states are also trying to regulate digital platforms (intermediaries) by imposing various rules and regulations, primarily relying on an older legal framework. Internationally, there is no universally recognized guideline to regulate content moderation. At the same time, there is no established principle to hold intermediaries liable in case of any violation of human rights. Moreover, because of their business model, intermediaries being pressurized by the states tend to take down more content than necessary, resulting in the suppression of free speech.

Intermediaries, while operating globally, often fail to understand the cultural context of the content. They struggle to balance diverse cultural variables and diverse sets of rules imposed by states. Facebook has recently established an oversight board to bring a cultural perspective to its content moderation system. The oversight board consists of experts from different parts of the world who represent various cultural backgrounds. From the analysis of some of the decisions by the Oversight Board, it is visible that while deciding a case, they mainly rely on the context of content such as user, location, language, audience etc. This Board is an attempt to accommodate and respect cultural diversity in their content moderation system. At the same time, this Board is proposing corrections to its community standards. This step taken by Facebook is praiseworthy in terms of its wiliness to respect human rights to the maximum possible extent. However, due to various limitations discussed throughout this thesis, it is challenging to protect freedom of expression for many users and audiences.[222]

There is no straightforward solution to this issue of governing freedom of expression on digital platforms. This is a complex legal issue, and it must be dealt with care. The UN bodies, states, regional organizations, intermediaries, civil society, think tanks, and other stakeholders have a role in this whole system. In the latest report on "Disinformation and Freedom of Opinion and Expression," the special rapporteur, "while acknowledging the complexities and challenges posed by disinformation in the digital age, finds that the responses by States and companies have been problematic, inadequate, and detrimental to human rights."[223] However, she did not distinguish digital platforms from our traditional understanding of media. As argued throughout this thesis, digital platforms are a particular type of media, and they must be treated accordingly. The mere recognition of digital platforms as "special media" might not change the

[222] If a user cannot exercise his right to freedom of expression, the audience might be deprived of his right to information and other associated rights.
[223] GA, Report of the Special Rapporteur on the Promotion and the Protection of the Right to Freedom of Opinion and Expression, 47th session, UN Doc. A/HRC/47/25, 13 April 2021.

current course (state-centric approach) of human rights law; however, it may pave the way for the further development of international human rights law.

To address the research questions of this thesis, I suggest having a universally accepted guideline for this type of "special media" platform where intermediaries' liability will be established with its scope, and there will also be scope for state intervention. Such guidelines will determine the extent of the self-governance of intermediaries and when states can intervene in their self-governance. An independent body (representing diverse cultures and expertise) may be appointed to oversee their self-governance, where cultural diversity will be respected and protected. If we can develop such a legal framework for digital platforms, we can express ourselves globally.

To conclude, the disruptive impacts of transitioning from Analog City to Digital City are unlikely to abate anytime soon. Before it seized on and helped spark the Protestant Reformation, the printing press had been around for 70 years. In comparison, the World Wide Web has only been around for almost 30 years, while Google, Facebook, and Twitter were established in 1998, 2004, and 2006, respectively. The digital era may still be in its early stages, with enormous changes to come.

As George Orwell puts it: “If large numbers of people are interested in freedom of speech, there will be freedom of speech, even if the law forbids it; if public opinion is sluggish, inconvenient minorities will be persecuted, even if laws exist to protect them.” Freedom of expression is still an experiment, and no one can predict the consequence of offering worldwide platforms to billions of individuals in the digital era. However, the experiment is noble and one that should be continued.

Bibliography

PRIMARY SOURCES: JURISPRUDENCE

Arslan v. Turkey, (Application no. 23462/94) Eur Ct HR (1999)

Delfi AS v Estonia, Eur. Ct. H.R. No. 64569/09, (16 June 2015)

Google France S.A.R.L. and Google, Inc. v. Louis Vuitton Malletier S.A. and Others, Joined Cases of C-236/08 to C-238/08, ECLI:EU:C:2010:159. Cour de Cassation [Final court of appeals] *Google France S.A.R.L. and Google, Inc. v. Louis Vuitton Malletier S.A. and Others,* Joined Cases of C-236/08 to C-238/08, 23 March 2010.

L'Oréal S.A. and Others v. eBay International A.G. and Others, Case C-324/09, ECLI:EU:C:2011:474.

Prosecutor v Brđanin, IT-99-36-T, Judgment, para 80 (ICTY, Sept. 1, 2004).

United States v. von Weizsaecker, Judgment (Intl Mil Trib Sept 30, 1946), *reprinted in* F.R.D. 161–163 (1946).

Verein gegen Tierfabriken Schweiz (VgT) v Switzerland [2001] 334 EHRR 159.

Von Hannover v Germany, [2005] 40 EHRR 1;

PRIMARY SOURCES: LEGISLATIONS

Communications Decency Act, 1996 47 U.S.C. §230

Ending Support for Internet Censorship Act, S.1914, 116th Cong. (2019)

The Computer Crimes Act 1997, Act 563, as amended on 1 January 2006,

The Computer Fraud and Abuse Act 1986, 18 USC

The Computer Misuse and Cybersecurity Act, Act 19 of 1993 (Revised 31 July 2007), as amended on 2 January 2011

UK Computer Misuse Act 1990,

PRIMARY SOURCES: INTERNATIONAL INSTRUMENTS (UN and EU)

GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 71st Sess, UN Doc. A/71/373, 6 September 2016.

GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 38th Sess, UN Doc. A/HRC/38/35, 6 April 2018.

GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 74th Sess, UN Doc. A/74/486, 9 October 2019.

GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 32nd Sess, UN Doc. A/HRC/32/38, 11 May 2016.

International Convention on the Elimination of All Forms of Racial Discrimination, 21 December 1965, 660 UNTS 195 (entered into force 4 January 1969)

General comment no. 34, Article 19, Freedoms of opinion and expression, UN Human Rights Committee (HRC), 102nd Session CCPR/C/GC/34, 12 September 2011

General Comment no. 11, Article 20, Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred, Human Rights Committee (HRC), 19th Session, CCPR/C/GC/11, 29 July 1983

United Nations Treaty Collection, Chapter IV, Human Rights, ICCPR, New York, 16 December 1976, online:

<https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATYandmtdsg_no=IV4andchapter=4andlang=en#ENdDec>.

Human Rights Committee: *J.R.T. and the W.G. Party v Canada*, Communication No 104/1981 (18 July 1981), UN Doc A/38/40 (Supp No. 40)

Human Rights Committee: *Malcolm Ross v Canada*, Communication No. 736/1997 (1 May 1996), UN Doc CCPR/C/70/D/736/1997,

Human Rights Committee: *Kasem Said Ahmad and Asmaa Abdol-Hamid v Denmark*, Communication No 1487/2006, UN Doc CCPR/C/92/D/1487/2006, 18 April 2008

Human Rights Council: *Incitement to Racial and Religious Hatred and the Promotion of Tolerance: Report of the High Commissioner for Human Rights*, 2nd Sess, UN Doc. A/HRC/2/6, 20 September 2006

Human Rights Council: Report of the United Nations High Commissioner for Human Rights and Follow-Up to the World Conference on Human rights, Addendum, Expert Seminar on the links between Article 19 and 20 of the International Covenant on Civil and Political Rights: "Freedom of expression and advocacy of religious hatred that constitutes incitement to discrimination, hostility or violence", 10th Sess, UN Doc. A/HRC/10/31/Add.3, 16 January 2009

GA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and, 7th Sess, UN Doc. A/67/357, 7 September 2012

United Nations Human Rights Office of the High Commissioner, "Guiding Principles on Business and Human Rights" (2011), online (pdf):

<https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf> 1 at p 5.

GA, Report of the Special Rapporteur on freedom of religion or belief, 25th Sess, UN Doc. A/HRC/25/58, 26 December 2013

Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance on manifestations of racism, racial discrimination, xenophobia and related intolerance, 26th Sess, UN Doc. A/HRC/26/49, 6 May 2014

Committee on the Elimination of Racial Discrimination: *L.K. v The Netherlands*, Communication No 4/1991, UN Doc CERD/C/42/D/4/1991, 16 March 1993

Committee on the Elimination of Racial Discrimination: *Quereshi v Denmark*, Communication No 33/2002, UN Doc CERD/C/66/D/33/2003, 10 March 2004

Nations, United, "United Nations Strategy and Plan of Action on Hate Speech," (May 2019), online (pdf): *United Nations* <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOP-SIS.pdf.>

Committee on the Elimination of Racial Discrimination: *L.R. v Slovak Republic*, Communication No 31/2003, UN Doc CERD/C/66/D/31/2003, 10 March 2005.

Human Rights Council, "Report of the independent international fact-finding mission on Myanmar", 12 September 2018, A/HRC/39/64.

High Commissioner for Human Rights, Opening Statement to the 36th session of the Human Rights Council, 11 September 2011.

Council of Europe, Convention for the Protection of Human Rights and Fundamental Freedoms, Rome, 4.XI.1950.

Official Records of the General Assembly, 64th Sess, Supplement No. 18 (A/64/18), annex VIII at para 13.

GA, Report of the Special Rapporteur on the Promotion and the Protection of the Right to Freedom of Opinion and Expression, 47th session, UN Doc. A/HRC/47/25, 13 April 2021

GA, Report of the Special Rapporteur on the situation of human rights in Myanmar, Advance Unedited Version, A/HRC/37/70, 9 March 2018.

Universal Declaration of Human Rights, 10 December 1948, G.A. Res. 217 A(III), U.N. Doc. A/810.

EC Convention on Cybercrime (adopted 23 November 2001, entered into force 1 July 2004) 185 European Treaty Series.

Explanatory Report of the Convention on Cybercrime (13 November 2001) CM (2001)144 addendum, https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804d873c4

Human Rights Council, UN General Assembly Resolution 16/18, 16th Sess, UN Doc. A/HRC/RES/16/18, 12 April 2011 (The Rabat Plan of Action).

International Covenant on Civil and Political Rights, 19 December 1966, 999 UNTS 171 arts 9—14 (entered into force 23 March 1976).

SECONDARY SOURCES: BOOKS

"Free Speech, Media Freedom and Regulation of Online Speech" in *Dimensions of Free Speech* Philosophy and Politics - Critical Explorations (Cham: Springer International Publishing, 2021) 93.

Article 19, "Hate Speech" Explained: A Toolkit (London, 2015)

Article 19, *Prohibiting Incitement to Discrimination, Hostility or Violence* (London, 2012)

Caplan, Robyn, *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*, by Robyn Caplan (Melbourne: Data & Society Research Institute, 2018).

Coliver, Sandra & Article 19 (Organization), eds, *The Article 19 freedom of expression handbook: international and comparative law, standards, and procedures* (London: Article 19, 1993).

Culnan, M.J, & M-L Markus, 'Information Technologies' in L. Putnam & D. Mumby (Eds.), Handbook of Organizational Communication (Beverly Hills, CA: Sage 1987) 420.

David Livingstone Smith, *Less than Human: Why we Demean, Enslave and Exterminate Others* (New York: St. Martin's Press, 2011).

DK, Citron, *Hate Crimes in Cyberspace*, (Harvard, MA: Harvard University Press, 2014)

Eliza, Varney, "Art.21 Freedom of Expression and Opinion, and Access to Information" in *UN Conv Rights Pers Disabil* (Oxford University Press, 2018).

Fung, Archon, Mary Graham & David Weil, *Full disclosure: the perils and promise of transparency* (New York: Cambridge Univ. Press, 2007) at 59-63.

Gillespie, Tarleton, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. (New Haven: Yale University Press, 2018).

Gillespie, Tarleton, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (New Haven: Yale University Press, 2018).

Gordon, Gregory S, Atrocity Speech Law: Foundation, Fragmentation, Fruition (Oxford: Oxford University Press, 2017).

Herz, Michael, and Peter Molnar, *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (Cambridge, UK: Cambridge University Press, 2012).

Hill, John Lawrence, *The Prophet of Modern Constitutional Liberalism: John Stuart Mill and the Supreme Court*, 1st ed (Cambridge University Press, 2020).

Jenkins, Henry, *Convergence Culture: Where Old and New Media Collide* (New York: New York University Press, 2006).

Marshall McLuhan, *Understanding Media: The Extensions of Man*, (New York City: McGraw-Hill, 1964) chapter 1.

Milo, Dario, Glenn Penfold & Anthony Stein, "Chapter 42 Freedom of Expression" 200.

Oppenheim, John, & Willem-Jan Van der Wolf, *Global War Crimes Tribunal Collection* (Nijmegen, The Netherlands: Global Law Association, 1997).

Schaumburg-Müller, Sten, "Private Life, Freedom of Expression and the Role of Transnational Digital Platforms: A European Perspective" in YSEC Yearbook of Socio-Economic Constitutions (Cham: Springer International Publishing, 2022).

Solove, Daniel J, *The Future of Reputation: Gossip, Rumor, and the Privacy on the Internet*, (London: Yale University Press, 2007)

Steuter, Erin, & Deborah Wills, *At War with Metaphor: Media, Propaganda, and Racism in the War on Terror* (Lanham: Lexington, 2009).

Wilson, Richard Ashby, *Incitement on Trial: Prosecuting International Speech Crimes*, (Cambridge: Cambridge University Press, 2017) in Cambridge Studies in Law and Society.

Zimbardo, Philip, "*The Lucifer Effect: Understanding How Good People Turn Evil*" (New York City: New York Random House, 2013).

Zuboff, Shoshana, *The Age of Surveillance Capitalism*, (London: Profile Books, 2019).

SECONDARY SOURCES: JOURNAL ARTICLES

Balkin, Jack M, "Digital Speech and Democratic Culture: A Theory of Freedom of Expression for The Information Society" (2004) 79:1 NYU L Rev 1–58.

Balkin, Jack M, "Free Speech is a Triangle" (2018) 118:7 Columbia L Rev 2011–2056.

Balkin, Jack M, "The Future of Free Expression in a Digital Age" (2009) 36:2 Pepp L Rev 427.

Boyle, Kevin, "Hate Speech— the United States Versus the Rest of the World?" (2001) 53 Maine L Rev 487

Brown, Alexander, "What is so special about online (as compared to offline) hate speech?" (2018) 18: 3 Ethnicities at 297.

Bunting, Mark, "From Editorial Obligation to Procedural Accountability: New Policy Approaches to Online Content in the Era of Information Intermediaries" (2018) 3:2 J Cyber Poly, online: <https://www.ssrn.com/abstract=3185005>.

Carol Pauli, Killing the Microphone: When Broadcast Freedom Should Yield to Genocide Prevention, (2010) 61: 4 Ala L Rev 665.

Citron, Danielle Keats, & Helen Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age" (2011) 91 BUL Rev 1435;

Dawkins, Robert, "Online Liberty: Freedom of Expression in the Information Age" (2001) 10 Dal J Leg Stud 102.

Dias Oliva, Thiago, "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression" (2020) 20:4 H R L Rev 607–640.

DK, Citron, and Norton H, “Intermediaries and hate speech: Fostering digital citizenship for our information age” (2011) 91: Boston U L Rev 1435.

Douek, Evelyn, “The Limits of International Law in Content Moderation” (2021) 6 UC Irvine J Intl Transnational & Compa L, online: <https://www.ssrn.com/abstract=3709566>.

Enarsson, Therese, & Simon Lindgren, “Free speech or hate speech? A legal analysis of the discourse about Roma on Twitter” (2018) Information & Communications Technology L at 4.

Estelles-Arolas, Enrique & Fernando Gonzales-Ladron-de-Guevara, “Towards an Integrated Crowdsourcing Definition”, (2012) 38 J Info Sci 189.

Finnemore, Martha, & Kathryn Sikkink, “International Norm Dynamics and Political Change” (1998) 52:4 Intl Organizations 887.

Goldman, Eric, “The Ten Most Important Section 230 Rulings” (2017) 20 Tul J Tech & Intell Prop 1.

Golia, Angelo Jr, “Beyond Oversight: Advancing Societal Constitutionalism in the Age of Surveillance Capitalism” (2021) SSRN Journal, online: <https://www.ssrn.com/abstract=3793219>.

Gordon, Gregory S., A War of Media, Words, Newspapers and Radio Stations’: The ICTR Media Trial Verdict and a New Chapter in the International Law of Hate Speech (2004) 45 VA J Intl L 139.

Grimmelmann, James, “The Virtues of Moderation” (2015) 17: 42 Yale J L & Tech 42

Helberger, Natali, Jo Pierson & Thomas Poell, “Governing online platforms: From contested to cooperative responsibility” (2018) 34:1 The Information Society 1–14.

Irving, Emma, “Suppressing Atrocity Speech on Social Media” (2019) 113 American Journal of International Law 256–261.

Klonick, Kate, “The New Governors: The People, Rules, and Processes Governing Online Speech” (2018) 131 Harv L Rev 1598–1670.

Koltay, András, “The Protection of Freedom of Expression from Social Media Platforms” (2022) 73:2 Mercer L Rev 523–589.

Kuczerawy, Aleksandra, “The Power of Positive Thinking: Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression” (2017) 3 J Intellectual Property Info Tech and Electronic Commerce L 182.

McGoldrick, David, & Therese O’Donnell, “Hate-speech laws: consistency with national and international human rights law” (1998) 18:4 Legal Studies 453.

McGoldrick, Dominic, & Therese O‘Donnel, “Hate-speech law: consistency with national and international human rights law” (1998) 18:4 Leg Studies 453

McKenna, Katelyn YA, and John A Bargh, Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology (2000) 4 Personality & Soc Psychology Rev 57.

Mulligan, Christina, “Technological Intermediaries and Freedom of the Press” (2013) 66 SMU L Rev 157

Myers West, Sarah, “Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms” (2018) 20:11 New Media & Society 4366–4383.

O’Neill, Michael Edmund, Old Crimes in New Bottles: Sanctioning Cybercrime, (2000) 9 Geo Mason L Rev 237.

Post, Robert C., “Racist Speech, Democracy, and the First Amendment” (1991) 32 William and Mary L Rev 267

Sander, Barrie, “FREEDOM OF EXPRESSION IN THE AGE OF ONLINE PLATFORMS”: (2020) 43:4 Fordham Intl L J 68.

Scanlon, Thomas, “A Theory of Freedom of Expression” (1972) 1:2 Phil & Pub Aff 204–226.

Suzor, Nicolas P, et al, “Human rights by design: The responsibilities of social media platforms to address gender-based violence online” (2018) 11:1 Policy & Internet 83

Tanenbaum, Robert S., “Preaching Terror: Free Speech or Wartime Incitement,” (2005) 55 American U L Rev 785

Tsesis, Alexander, “Terrorist Speech on Social Media”, (2017) 70 Vand L Rev 651.

Viljoen, Frans. “Inciting violence and propagating hate through the Media: Rwanda and the limits of international criminal law” (January 2005) 26:1 Obiter 58.

Walther, Joseph B., “Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction” (1996) 23: 1 Communication Research 3.

Wenguang, Yu, “Internet Intermediaries’ Liability for Online Illegal Hate Speech” (2018) 13:3 Frontiers of L in China 342

Wu, Felix T., “Collateral Censorship and the Limits of Intermediary Immunity” (2011) 87 Notre Dame L Rev 293;

SECONDARY SOURCES: REPORTS

Business for Social Responsibility, “Human Rights Impact Assessment: Facebook in Myanmar” (2018) at 12, online (pdf):< https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf >.

———, “Facebook’s Content Moderation Rules Are a Mess | Brennan Center for Justice”, (22 February 2021), online: *The Brennan Center for Justice* <https://www.brennancenter.org/our-work/analysis-opinion/facebooks-content-moderation-rules-are-mess>.

“Community Standards Enforcement | Transparency Center”, (8 February 2022), online: *Meta* <https://transparency.fb.com/data/community-standards-enforcement/>.

“DEPARTMENT OF JUSTICE’S REVIEW OF SECTION 230 OF THE COMMUNICATIONS DECENCY ACT OF 1996”, (3 June 2020), online: <https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996>.

“Detecting violations | Transparency Center”, (10 February 2022), online: *Meta* <https://transparency.fb.com/enforcement/detecting-violations/>.

“Hate Speech | Transparency Center”, (8 February 2022), online: *Meta* <https://transparency.fb.com/policies/community-standards/hate-speech/>.

“Hate speech policy - YouTube Help”, (8 February 2022), online: *YouTube Help* <https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436>.

“Innocence of Muslims Controversy”, *Berkley Center for Religion, Peace & World Affairs* (2013), http://berkleycenter.georgetown.edu/essays/em-innocence-ofmuslims-em-controversy>.

“Meet the Board | Oversight Board”, online: <https://oversightboard.com/meet-the-board/>.

“Oversight Board | Independent Judgment. Transparency. Legitimacy.”, online: <https://oversightboard.com/>.

"The Twitter rules: safety, privacy, authenticity, and more", (8 February 2022), online: *Twitter Help Center* <https://help.twitter.com/en/rules-and-policies/twitter-rules>.

AfricaNews, Redaction, "Rwanda silencing YouTubers with 'abusive' legal framework - Human Rights Watch", (16 March 2022), online: *Africanews* <https://www.africanews.com/2022/03/16/rwanda-silencing-youtubers-with-abusive-legal-framework-human-rights-watch/>.

Angwin, Julia & Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children", (28 June 2017), online: *ProPublica* <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms?token=sDXc9-_dthGWtDlWPbCrrxBIUqQRIggX>.

Bartow, Ann, "Section 230 Keeps Platforms for Defamation and Threats Highly Profitable", (10 November 2017), online: *The Recorder* <https://www.law.com/therecorder/sites/therecorder/2017/11/10/section-230-keeps-platforms-for-defamation-and-threats-highly-profitable/>.

Bennett, Shea, 10 Easy Ways to Get More Retweets on #Twitter, (5 January 2015), online: *Adweek*, <www.adweek.com/digital/get-more-retweets-twitter/>.

Brandom, Russell, "New rules challenge Google and Facebook to change the way they moderate users", (7 May 2018), online: *The Verge* <https://www.theverge.com/2018/5/7/17328764/santa-clara-principles-platform-moderation-ban-google-facebook-twitter>.

Columbia University, Global Freedom of Expression, "*Delfi v Estonia*", online: <https://globalfreedomofexpression.columbia.edu/cases/delfi-as-v-estonia/>.

Dawson, Aimee, “Facebook censors 30,000 year-old Venus of Willendorf as ‘pornographic’”, (27 February 2018), online: *The Art Newspaper* <https://www.theartnewspaper.com/2018/02/27/facebook-censors-30000-year-old-venus-of-willendorf-as-pornographic>.

Dias, Talita de Souza, "Propaganda and Accountability for International Crimes in the Age of Social Media: Revisiting Accomplice Liability in International Criminal Law" (04 April, 2018) *Opinio Juris*, online: <http://opiniojuris.org/2018/04/04/propaganda-and-accountability-for-international-crimes-in-the-age-of-socialmedia-revisiting-accomplice-liability-in-international-criminal-law/>.

Donovan, Joan, “Why social media can’t keep moderating content in the shadows”, (6 November 2020), online: *MIT Technology Review* <https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/>.

Douek, Evelyn, “More Content Moderation Is Not Always Better” *Wired* (2 June 2021), online: <https://www.wired.com/story/more-content-moderation-not-always-better/>.

Egbunike, Nwachukwu & Kofi Yeboah, “Twitter’s deal with Nigerian government sacrifices digital rights”, (17 January 2022), online: *Global Voices* <https://globalvoices.org/2022/01/17/twitters-deal-with-nigerian-government-sacrifices-digital-rights/>.

Etter, Lauren, "What Happens When the Government Uses Facebook as a Weapon?" (7 December 2017), online: *Bloomberg Businessweek* < https://www.bloomberg.com/news/features/2017-12-07/how-rodrigo-duterteturned-facebook-into-a-weapon-with-a-little-help-from-facebook>

Frayer, Lauren, “How the Spread of Fake Stories in India Has Led to Violence” (July 17, 2018), online: *NPR* <https://www.npr.org/2018/07/17/629896525/how-the-spread-of-fake-stories-in-india-has-led-to-violence>;

Friedman, Uri, “An American in ISIS’s Retweet Army”, *Atlantic* (Aug. 29, 2014), online: <http://www.theatlantic.com/international/archive/2014/08/anamerican-in-isis-retweet-army/379208/>

Gilsinan, Kathy, “Is ISIS’s Social-Media Power Exaggerated?”, (Feb 23, 2015), online: *Atlantic* <http://www.theatlantic.com/international/archive/2015/02/is-isisssocial-media-power-exaggerated/385726/>

Golemanova, Ralitsa, “What Is Content Moderation? | Types of Moderation & Content to Moderate”, (8 September 2021), online: *Imagga Blog* <https://imagga.com/blog/what-is-content-moderation/>.

Harper, Cindy, “Facebook censorship hinders Cuban journalists reporting on corruption”, (28 November 2021), online: *Reclaim The Net* <https://reclaimthenet.org/facebook-censorship-hinders-cuban-journalists-reporting-on-corruption/>.

Ilori, Tomuwa, “In Nigeria, the government weaponises the law against online expression”, (17 December 2021), online: *Global Voices* <https://globalvoices.org/2021/12/17/in-nigeria-the-government-weaponises-the-law-against-online-expression/>.

Ingram, Mathew, "Facebook now linked to violence in the Philippines, Libya, Germany, Myanmar, and India" (5 September 2018), online: *Columbia Journalism Review* <https://www.cjr.org/the_media_today/facebook-linked-to-violence.php>.

Ingram, Mathew, “In some countries, fake news on Facebook is a matter of life and death” (21 November 2017) *Columbia Journalism Review*, online :< https://www.cjr.org/analysis/facebook-rohingya-myanmar-fake-news.php >.

Kantor, Arcadiy, “Measuring Our Progress Combating Hate Speech”, (19 November 2020), online: *Meta* <https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/>.

Kilovaty, Ido, “Are Tech Companies Becoming the Primary Legislators in International Cyberspace?” (28 March 2019) Lawfare, online:<https://www.lawfareblog.com/are-techcompaniesbecomingprimarylegislators-internationalcyberspace?fbclid=IwAR1T9o2T1KQn-RQWYgRY_pIyjXiByy1-Aw_aPrMkXrrf3Nz6uhbH15HhTxw#__prclt=pLTEhkPm>.

Kuhn, Anthony, “Activists In Myanmar Say Facebook Needs To Do More To Quell Hate Speech”, *NPR* (14 June 2018), online: <https://www.npr.org/2018/06/14/619488792/activists-in-myanmar-say-facebook-needs-to-do-more-to-quell-hate-speech>.

Loh, Alicia, “Google v Equustek: United States Federal Court Declares Canadian Court Order Unenforceable” (16 November 2017), online: <https://jolt.law.harvard.edu/digest/google-v-equustek-united-statesfederalcourt-declares-canadian-court-order-unenforceable>.

Lynch, Justin, "In South Sudan, Fake News Has Deadly Consequences" (09 June, 2017), online: *The Slate* <http://www.slate.com/articles/technology/future_tense/2017/06/in_south_sudan_fake_news_has_deadly_consequences.html>.

McLaughlin, Timothy, “How Facebook’s Rise Fueled Chaos and Confusion in Myanmar” *Wired* (6 July 2018), online: <https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/>.

Patel, Faiza & Laura Hecht-Felella, “Evaluating Facebook’s New Oversight Board for Content Moderation”, (19 November 2019), online: *Just Security* <https://www.justsecurity.org/67290/evaluating-facebooks-new-oversight-board-for-content-moderation/>.

Pozin, Ilya, 6 Qualities to Make Your Videos Go Viral, (7 August 2014) online: *Forbes* <www.forbes.com/sites/ilyapozin/2014/08/07/6-qualities-to-make-yourvideos-go-viral>.

Preez, Derek de, “UK’s Online Safety Bill - not robust enough to to tackle illegal content, nor does it protect freedom of expression”, (24 January 2022), online: *Diginomica* <https://diginomica.com/uks-online-safety-bill-not-robust-enough-tackle-illegal-content-nor-does-it-protect-freedom>.

Ruane, Kathleen Ann, “How Broad A Shield? A Brief Overview of Section 230 of the Communications Decency Act”, (2018), online: *Congressional Research Service*.

Safety, Twitter, “Updating our rules against hateful conduct”, (13 December 2021), online: *Twitter Safety* <https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate>.

Smedt, Tom De, Guy de Pauw, & Pieter Van Ostaeyen, “Automatic Detection of Online Jihadist Hate Speech” (February 2018) CTRS-007

Su, Sara, “Update on Myanmar” (15 August, 2018), online: *Facebook Newsroom* <https://newsroom.fb.com/news/2018/08/update-on-myanmar/>.

Team, The YouTube, “Our ongoing work to tackle hate”, (5 June 2019), online: *blog.youtube* <https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/>.

Thompson, Alan. *The Media and the Rwanda Genocide*, (London/Ottawa: Pluto Press/International Development Research Centre, 2007) International Development Research Centre.

SECONDARY SOURCES: NEWSPAPER NEWS AND ARTICLES

“Burning Cameroon: Images you're not meant to see” (25 June 2018) *BBC News* (Prime Minister Philemon Yang has blamed Cameroonians living abroad for using social media to “spread hate speech and terror” and “order murders”), online: < https://www.bbc.com/news/world-africa-44561929>.

“Q&A: Anti-Islam film”, *BBC News* (20 September 2012), online: <https://www.bbc.com/news/world-middle-east-19606155>.

“Social media rumors trigger violence in India; 3 killed by mobs” (May 25, 2018), online: *NBC News* <https://www.nbcnews.com/news/world/social-media-rumors-trigger-violence-india-3-killed-mobs-n877401>.

“Twitter receives record number of gov’t requests to remove posts”, *Al jazeera* (26 January 2022), online: <https://www.aljazeera.com/news/2022/1/26/twitter-sees-record-number-of-govt-demands-to-remove-content>.

Balakrishnan, Anita, “Facebook pledges to double its 10,000-person safety and security staff by end of 2018”, (31 October 2017), online: *CNBC* <https://www.cnbc.com/2017/10/31/facebook-senate-testimony-doubling-security-group-to-20000-in-2018.html>.

Beinart, Peter, “What Does Obama Really Mean by ‘Violent Extremism’?”, (Feb 20, 2015) *Atlantic*, online: <http://www.theatlantic.com/international/archive/2015/02/obamaviolent-extremism-radical-islam/385700/>

Bovard, Rachel, “How many times must Facebook be caught censoring the truth?”, *New York Post* (22 November 2021), online: <https://nypost.com/2021/11/22/how-many-times-must-facebook-be-caught-censoring-the-truth/>.

Eric Schmitt, “U.S. Intensifies Effort to Blunt ISIS’ Messages”, *New York Times* (Feb 16, 2015), online: <http://www.nytimes.com/2015/02/17/world/middleeast/us-intensifieseffort-to-blunt-isis-message.html>

Giusti, Kathy, “The Real Ice Bucket Challenge”, online: *Time* <https://time.com/3204261/the-real-ice-bucket-challenge/>.

Goel, Vindu, "In India, Facebook’s WhatsApp Plays Central Role in Elections" *The New York Times* (14 May 2018), online: <www.nytimes.com/2018/05/14/technology/whatsapp-india-elections.html>.

Hubbard, Ben, “Jihadists and Supporters Take to Social Media to Praise Attack on Charlie Hebdo”, *New York Times* (Jan. 11, 2015), online: <http://www.nytimes.com/2015/01/11/world/europe/islamicextremists-take-to-social-media-to-praise-charlie-hebdoattack> [http://perma.cc/CH92-LH5S]

Koetsier, John, “Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day”, (9 June 2020), online: *Forbes* <https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/>.

Maltby, Kate, “The Online Harms Bill is a threat to freedom of expression, but it’s buried in Government chaos”, (28 January 2022), online: *inews.co.uk* <https://inews.co.uk/opinion/online-harms-bill-threat-freedom-expression-buried-government-chaos-1427925>.

Moore, Jina, “Cambridge Analytica Had a Role in Kenya Election, Too” (20 March, 2018) *The New York Times*, online: https://www.nytimes.com/2018/03/20/world/africa/kenya-cambridge-analytica-election.html

Mozur, Paul, "A Genocide Incited on Facebook, With Posts from Myanmar’s Military"(Oct. 15, 2018) *The New York Times*, online:< https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>.

Roose, Kevin, Political Donors Put Their Money Where the Memes Are, *New York Times* (7 August 2017), online <www.nytimes.com/2017/08/06/business/media/politicaldonors-put-their-money-where-the-memes-are.html>.

Safi, Michael, "Sri Lanka accuses Facebook over hate speech after deadly riots" (14 March, 2018) *The Guardian*, online:< https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hatespeech>.

Sayare, Scott & Nicola Clark, “French Newspaper Publishes Cartoons Mocking Muhammad - The New York Times”, *The New York Times* (19 September 2012), online: <https://www.nytimes.com/2012/09/20/world/europe/french-magazine-publishes-cartoons-mocking-muhammad.html>.

Schlein, Lisa, "Hate Speech on Social Media Inflaming Divisions in CAR" (02 June 2018), online: *VOA News* <https://www.voanews.com/a/hate-speech-on-social-media-is-inflaming-divisions-in-central-africanrepublic/4420555.html>.

Schoolov, Katie, "Why content moderation costs billions and is so tricky for Facebook, Twitter, YouTube and others", (27 February 2021), online: *CNBC* <https://www.cnbc.com/2021/02/27/content-moderation-on-social-media.html>.

Scott Neuman, "Homeland Security Chief: Threat to U.S. Malls 'A New Phase' For Terrorists", *NPR* (Feb. 22, 2015), online: <http://www.npr.org/blogs/thetwoway/2015/02/22/388242488/homeland-security-chiefthreat-

Smith, Adam, "Facebook comments like 'white men are stupid' were algorithmically rated as bad as antisemitic or racist slurs", (4 December 2020), online: *The Independent* <https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-comments-algorithm-racism-b1766209.html>.

Stecklow, Steve, "Why Facebook is losing the war on hate speech in Myanmar" (15 August 2018) *Reuters,* online: < https://www.reuters.com/investigates/specialreport/myanmar-facebook-hate/ >.

Stecklow, Steve, "Why Facebook is losing the war on hate speech in Myanmar", *Reuters* (15 August 2018), online: <https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/>.

Suciu, Peter, "Trolls Continue To Be A Problem On Social Media", online: *Forbes* <https://www.forbes.com/sites/petersuciu/2020/06/04/trolls-continue-to-be-a-problem-on-social-media/>.

Thomas, Kris, "6 Deaths, 450 Arrests and Mass Protests. It Started With a Facebook Post.", *VICE* (20 October 2021), online: <https://www.vice.com/en/article/y3vmpb/bangladesh-violence-facebook-post>.

to-u-s-mallsa-new-phase-for-terrorists>.

Tracy, Thomas, "ISIS has mastered social media, recruiting 'lone wolf' attacks to target Times Square: Bratton", *NY Daily News*, online: <http://www.nydailynews.com/new-york/isis-recruiting-lone-wolfterrorists-targettimes-square-bratton-article-1.1941687>

Yurieff, Kaya, "Facebook's 'supreme court' just ruled against Facebook", *CNN* (28 January 2021), online: <https://www.cnn.com/2021/01/28/tech/facebook-oversight-board-first-decisions/index.html>.

SECONDARY SOURCES: MISCELLANEOUS

Benge, Hayden, *Who's liable? The Intersection of Free Speech and Content Regulation on Social Media Platforms* (Honors Thesis, UniversityofMississippi, 2019) [unpublished].

Gesetz zur Verbesserung der Rechtsdurchsetzung in den sozialen Netzwerken [Act to Improve Enforcement of Law in the Social Networks], BGBl. I, S. 3352 of Sept. 1, 2017

Haupt, Claudia, "Online Speech Regulation: A Comparative Perspective," Presented at the American Political Science Association, August (2018).

JM Berger & Jonathon Morgan, "The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter" (2015) online (pdf): < https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf>

Jona, Adalheidur Palmadottir & Iuliana Kalenikova, "Hate speech; an overview and recommendations for combating it" (2018) *Icelandic Human Rights Centre* online (pdf): <http://www.humanrights.is/static/files/Skyrslur/Hatursraeda/hatursraedautdrattur.pdf>

Klausen, Jytte, et al., “The YouTube Jihadists: A Social Network Analysis of Al-Muhajiroun’s Propaganda Campaign,” Perspectives on Terrorism 6, no. 1 (2012).

Muller, Karsten, and Carlo Schwarz, “Fanning the Flames of Hate: Social Media and Hate Crime” Working Paper Series, (2018) at 2, *University of Warwich* online (pdf): <https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/373-2018_schwarz.pdf>.

Roberts, Sarah T., Behind The Screen: The Hidden Digital Labor of Commercial Content Moderation, (2014) Phd Thesis, University of Illinois, Chicago, IL